PROFILED REGULATORY SITES USEFUL FOR GENE CONTROL

1. FIELD OF THE INVENTION

The invention relates to methods for quantitative profiling of chromatin sensitivity to a DNA modifying agent. The invention also relates to methods for identifying regulatory sites in a genomic locus and to methods for determining chromatin architecture in a genomic locus. The invention also relates to the use of profiled regulatory sites, databases comprising the same, and their use in regulating gene expression, disease diagnosis and therapy, and identification of therapeutic drugs.

2. BACKGROUND OF THE INVENTION

Understanding the human genome requires comprehensive identification of DNA elements that are functional *in vivo*. A major class of such sequences are those which have a role in regulating genomic activity. Regulatory factors interact with chromatin in a site-specific fashion to bring the genome to life. All genes are controlled at multiple levels through the interaction of regulatory factors with gene-proximal or, in some cases, distant *cis*-regulatory sites. The nucleoprotein complexes formed by such interactions may be tissue or developmental stage-specific, or they may be constitutive, depending on the regulatory requirements of their cognate gene. While our knowledge of the patterns of gene expression in diverse tissues and under a wide-ranging set of conditions has grown substantially in recent years, this growth has not been paralleled by a comparable increase in our knowledge of regulatory factors that control specific genes affecting specific cellular or disease processes.

The basic chromatin fiber consists of an array of nucleosomes, each packaging around 200 base pairs of DNA; 146 is wound around the histone octamer, with the remainder forming a link to the next nucleosome. In eukaryotic cells, all genomic DNA in the nucleus is packaged into chromatin, the architecture of which plays a central role in regulating gene expression (for reviews see Felsenfeld, G. & Groudine, M., 2003, *Nature* 421, 448-53; Felsenfeld, G., 1992, *Nature* 355, 219-24; Brownell, J. E. & Allis, C. D., 1996, *Curr Opin Genet Dev* 6, 176-84; Kingston, R. E., Bunker,

C. A. & Imbalzano, A. N., 1996, *Genes Dev* 10, 905-20; Tsukiyama, T. & Wu, C., 1997, *Curr Opin Genet Dev* 7, 182-91; Wolffe, A. P., Wong, J. & Pruss, D., 1997, *Genes Cells* 2, 291-302; Kadonaga, J. T., 1998, *Cell* 92, 307-13; Struhl, K., 2001, *Science* 293:1054-1055). At a global level, this packaging serves two purposes: (i) it is physically necessary to condense the mass of sequence information into a well-ordered regular structure that can be contained within the nucleus; and (ii) it imparts a level of site-specific 'epigenomic' information (Felsenfeld, G., 1992, *Nature* 355, 219-24), for example discriminating between sequences which are never to be transcribed and are stored in highly condensed heterochromatin, and those sequences which are actively transcribed and are maintained in a more accessible chromatin state.

Gene expression is regulated by several different classes of *cis*-regulatory DNA sequences including enhancers, silencers, insulators, and core promoters (Felsenfeld and Groudine, 2003, *Nature* 421, 448-53; Butler and Kadonga, 2002, *Genes Dev* 16: 2583-2592; Gill, G., 2001, *Essays Biochem* 37: 33-43). The core promoter is the site of formation of the RNA pol II transcription complex. Enhancers and silencers act over distances of several kilobases (or more) to potentiate or silence pol II function. Insulator sequences prevent enhancers and silencers targeted to one gene from inappropriately regulating a neighboring gene. Larger more complex elements comprising multiple enhancer and/or silencers have come to light which coordinate the activity of linked genes over large chromosomal domains ('Locus Control Regions' or 'Domain Control Regions') (reviewed in Li *et al.*, 2002, Blood 100, 3077-86; Hardison, R.C., 2001, *Proc Natl Acad Sci U S A* 98:1327-1329). Activation of *cis*-regulatory elements in the context of chromatin requires the cooperative binding of regulatory factors (Felsenfeld, G., 1996, *Cell* 86, 13-9). This active state is most commonly addressed by measuring the sensitivity of the underlying DNA sequences to digestion with nucleases (e.g., DNaseI) in the context of chromatin (Weintraub, H. & Groudine, M., 1976, *Science* 193, 848-56; Elgin, S. C., 1981, *Cell* 27, 413-5). Multiprotein complexes exist in cells that allow specific destabilization of nucleosomes at promoters, facilitating the binding of sequence-specific factors and the general transcriptional machinery (Kingston, R. E., Bunker, C. A. & Imbalzano, A. N., 1996, *Genes Dev* 10, 905-20; Svaren, J., Horz, W., 1996, *Curr Opin Genet Dev.* 6:164-170; Tsukiyama, T. & Wu, C., 1997, *Curr Opin Genet Dev* 7, 182-91). Posttranscriptional modifications of chromatin components,

particularly histone RStylation, play important roles in regulating chromatin structure and gene activity (Brownell, J. E. & Allis, C. D., 1996, *Curr Opin Genet Dev* 6, 176-84; Grunstein, M., 1997, *Nature*. 389:349-352; Wolffe, A. P., Wong, J. & Pruss, D., 1997, *Genes Cells* 2, 291-302; Kadonaga, J. T., 1998, *Cell* 92, 307-13; Struhl, K., 1998, *Genes Dev* 12, 599-606).

Activation of tissue-specific genes during development and differentiation occurs first at the level of chromatin accessibility and results in the formation of transcriptionally-competent genetic loci characterized by increased sensitivity (relative to inactive loci) to digestion with DnaseI (Groudine *et al.*, 1983, *Proc Natl Acad Sci U S A*. 80:7551-7555; Tuan *et al.*, 1985, *Proc Natl Acad Sci U S A*. 82:6384-6388; Forrester *et al.*, 1986, *Proc Natl Acad Sci U S A*. 83:1359-1363). Loci in an accessible chromatin configuration can subsequently respond to acutely activating signals, often conveyed by non-tissue-specific transcriptional factors that can gain access to the open locus and recruit or activate the basal transcriptional machinery.

The initial observation that active genes reside within domains of generally increased sensitivity to nucleases was made nearly 30 years ago (Weintraub, H. & Groudine, M., 1976, *Science* 193, 848-56). Since this time, such data had been accumulated for a number of human gene loci (Pullner *et al.*, 1996, *J Biol Chem* 271: 31452-31457) and those in other vertebrates (Koropatnick and Duereksen, 1987, *Dev Biol* 122: 1-10; Stratling *et al.*, 1986, *Biochemistry* 25: 495-502). The chromatin domain phenomenon is particularly striking in *Drosophila*, where distinct transitions between DNase-sensitive and DNase-resistant chromatin can be documented (Farkas *et al.*, 2000, *Gene* 253: 117-136).

Focal alterations in chromatin structure are the hallmark of active regulatory sequences in eukaryotic genomes. The literature connecting DNaseI-hypersensitive sites with genomic regulatory elements is extensive. DNase hypersensitivity studies had been employed to delineate the transcriptional regulatory elements of over 100 human gene loci. Typically, between 1 and 5 hypersensitive sites had been visualized for each of these loci. However, only a fraction of these had been precisely localized at the sequence level.

A critical defining feature of HSs is that the function of the DNA sequence component, i.e. its complex-forming activity, is intrinsic. The principal evidence for this is the fact that these sequences can be excised and inserted into other positions in the genome, where they exhibit the same functional chromatin activities. Substantial

experimental experience from model systems has revealed that HSs can form when included in either constructs used to create stably transfected cell lines (Fraser *et al.*, 1990 *Nucleic Acids Res* 18:3503-3508)or transgenic animals (Lowrey *et al.*, 1992, *Proc Natl Acad Sci U S A* 89, 1143-7; Levy-Wilson *et al.*, 2000, *Mol Cell Biol Res Commun* 4, 206-11).

An important finding has been that HS sequences are rendered functional only upon assembly into nuclear genomic chromatin. These DNA sequences are thought to potentiate formation of a nucleoprotein complex in a manner that dramatically increases its probability of activation vs. neighboring DNA regions. They are hypothesized to adopt a particular topological confirmation, which lowers the free energy for coalescence of a limited set of proteins, some in contact with DNA, and some in contact only with another protein in the complex. This results in the formation of a nucleoprotein complex which is pre*cis*ely correlated with a particular sequence. The formation of this complex takes plRS in an 'all-or-none' fashion (e.g., Felsenfeld *et al.*, 1996, *Cell* 86, 13-9; Boyes & Felsenfeld, 1996, *EMBO J* 15:2496-2507). The stochasticity of nucleoprotein complex formation can be manipulated through the introduction of point mutations or small deletions or insertions in critical DNA binding bases or in juxtaposed sequences that affect overall stability (e.g., Stamatoyannopoulos *et al.*, 1995, *EMBO J* 14, 106-16).

Cooperative binding of regulatory factors in the context of chromatin results in sequence-specific 'remodeling' of the local chromatin architecture (Felsenfeld and Groudine, 2003. *Nature* 421; 448-453). This focal 'remodeling' is the signature of active regulatory foci within genomic sequences and is detectable experimentally on the basis of pronounced sensitivity to cleavage when intact nuclei are exposed to DNA modifying agents, canonically the non-specific endonuclease DnaseI (Gross and Garrard 1988. *Annu. Rev. Biochem.* 57; 159-197, Elgin 1984. *Nature* 309; 213-4, Wu 1980. *Nature* 286; 854-860). The co-localization of DNaseI Hypersensitive Sites (HSs) with *cis*-active elements spans the spectrum of known transcriptional and chromosomal regulatory activities including transcriptional enhancers, promoters, and silencers, insulators, locus control regions, and domain boundary elements (Felsenfeld 1996. *Cell* 86, 13-9, Gross and Garrard 1988. *Annu. Rev. Biochem.* 57; 159-197, Burgess-Beusse *et al.*, 2002. *Proc. Natl. Acad. Sci. USA* 99; 16433-7). HSs have also been observed to coincide with sequences governing fundamental genomic processes

including attachment to the nuclear matrix (Jarman and Higgs 1988. *EMBO J.* 7; 3337-44, Kieffer *et al.,* 2002. *J. Immunol.* 168; 3915-3922), and recombination (Zhang *et al.*, 2002. *Proc. Natl. Acad. Sci USA* 99; 3070-3075), though their association with these lower level chromosomal processes is less easy to document owing to their ephemeral nature or cell-cycle specific appearance.

Property	*Definition*	*Examples*	*Reference*
Promoter	Transcriptional promoter	c-myc TBP Interleukin-6	Pullner *et al.*, 1996. *J. Biol. Chem.* 271; 31452-31457 Harland *et al.* 1992. *Genomics* 79; 479-482. Armenante *et al.*, 1999. *Nucl. Acids Res.* 27; 4483-4490.
Transcriptional Enhancer	Up-regulates transcription from linked gene	Beta-globin HS2 *apoB* enhancer CD34 enhancer	Kong *et al.*, 1997. *Mol. Cell Biol.* 17; 3959-65. Levy-Wilson *et al.*, 2000. *Mol. Cell Biol. Res. Commun.* 4; 206-211. Radomska *et al.*, 1998. *Gene* 222; 305-318.
Insulator	Demarcates gene regulatory domains	Beta-globin HS5 H19/Igf2	Li and Stamatoyannopoulos 1994. *Blood* 84; 1399-1401. Jones *et al.*, 2001. *Hum. Mol. Genet.* 10; 807-814.
Locus Control Region	Determines long-range chromatin structure and control of multiple linked genes	Beta-globin CD2 Adenine Deaminase	Grosveld 1999. *Curr. Opin. Genet. Dev.* 9; 152-157. Festentein *et al.*, 1996. *Science* 271; 1123-5. Aronow *et al.*, 1992. *Mol Cell. Biol.* 12; 4170-4185.
Transcriptional Silencer	Down-regulates transcription from linked gene	GATA3 silencer	Gregoire and Romeo, 1999. *J. Biol. Chem.* 274; 6567-6578.
Matrix Attachment Region	Tether chromatin to protein backbone	CD8 gene comples MARs	Kieffer *et al.*, 2002. *J. Immunol.* 168; 3915-22.
Origin of Replication (ORI)	Origin of DNA replication	*Puff II/9A* ORI	Urnov *et al.*, 2002. *Chromosoma* 11; 291-303.
Recombination Sites	Sites of frequent chromosome recombination or translocation	AML1/RUNX1 breakpoints in t(8;21) leukemia	Zhang *et al.*, 2002. *Proc. Natl. Acad. Sci. USA.* 99; 3070-3075.

DNase hypersensitivity studies collectively comprise the most successful and extensively validated methodology for discovery of regulatory sequences *in vivo*, and

had been employed to delineate the transcriptional regulatory elements of >100 human gene loci. Over 25 years of experimentation and legion publications by many investigators have established an inviolable connection between sites of DNase hypersensitivity *in vivo* and functional non-coding sequences that regulate the genome. In essentially every case where a major DNase HS has been adequately studied, a genomic regulatory activity has ultimately been disclosed, even if such function is not immediately apparent due to temporal or spacial restriction of activity (*e.g.*, Wai *et al.*, 2003. *EMBO J.* 22; 4489-4500). This is not merely a phenomenon of negative publication bias: since DNaseI HSs are biological phenomena of independent significance, they are extensively reported even without specific studies of their contribution to transcription. Conversely, in every published case where a regulatory sequence with documented in vivo activity (e.g., a promoter or enhancer discovered with other means) has been assayed for nuclease hypersensitivity, the expected result has been found.

It is now generally accepted that DNase HSs mark genomic sequences that bind regulatory factors *in vivo* with consequent disruption of the nucleosome array (Felsenfeld 1996. *Cell* 86; 13-19). Nuclease hypersensitive sites are biologically bounded by (a) the positions of flanking nucleosomes and (b) limits on the area of DNA over which thermodynamically stable nucleoprotein complexes may form. The extent of the regulatory domain is contained within the inter-nucleosomal interval, approximately 150-250bp. This interval corresponds to the size of sequence that is needed to place a canonical nucleosome and it has been a common assumption that HSs represent a break in the nucleosomal array that defines the vast majority of chromatin. A core domain can be identified which is restricted to a region of approximately 80-120 base pairs in length, over which critical DNA-protein interactions take place (e.g., Lowrey *et al.*, 1992. *Proc. Natl. Acad. Sci. USA* 89; 1143-1147). Cooperative binding of transcription factors to such core regions is sufficient to exclude a nucleosome *in vitro* (Adams and Workman, 1995. *Mol. Cell Biol.* 15; 1405-1421) and this is now accepted as a common mechanism for how these sites form *in vivo* (Boyes and Felsenfeld, 1996. *EMBO J.* 15; 2496-2507; Wallrath *et al.*, 1994. *Bioessays* 16; 165-170; Struhl, 2001. *Science* 293; 1054-1055).

In summary, DNase HSs are extensively validated markers of sequence-specific *in vivo* functionality and should therefore be presumed to be involved in regulation of neighboring genes until proven otherwise (Urnov 2003. *J. Cell Biochem.*

88; 684-694). *DNaseI hypersensitivity studies thus represent a powerful,* in vivo *approach to detection and analysis of biologically active sequences.*

Nuclease hypersensitive sites are biologically bounded by (1) the positions of flanking nucleosomes and (2) limits on the area of DNA over which thermodynamically stable nucleoprotein complexes may form. The extent of the regulatory domain is contained within the inter-nucleosomal interval, approximately 150-250bp. This interval corresponds to the size of sequence that is needed to plRS a canonical nucleosome and it has been a common assumption that HSs represent a break in the nucleosomal array that constitutes the vast majority of chromatin.

A core domain can be identified which is restricted to a region of approximately 80-120 base pairs in length, over which DNA-protein interactions take plRS (e.g., Lowrey *et al.*, 1992, *Proc Natl Acad Sci U S A* 89, 1143-7). Cooperative binding of transcription factors to such core regions is sufficient to exclude a nucleosome *in vitro* (Adams and Workman, 1995, *Mol Cell Biol* 15, 1405-1421) and this has been proposed as a common mechanism for how these sites may form *in vivo*. Nucleosomal mapping experiments have shown that HSs such as the *Drosophila hsp26* promoter (Lu *et al.*, 1995 *EMBO J.* 2; 4738-46) and the human β-globin HS2 (Kim and Murray, 2001, . *Int J Biochem Cell Biol* 33, 1183-92) are non-nucleosomal. It is thought that most HSs are non-nucleosomal in nature (Boyes and Felsenfeld, 1996, *EMBO J* 15:2496-2507; Wallrath *et al.*, 1994, *Bioessays* 16:165-170). These conclusions are well-supported in the Litreature (e.g., Struhl, 2001, *Science* 293:1054-1055). However several HSs are known to still have histone proteins and transcription factors, suggesting that HSs may exist in conjunction with a modified or partial nucleosome.

Flanking sequences surrounding the core region appear to modulate the activity of this core region, though this effect tapers off sharply. The boundaries of the sequences needed for hypersensitivity can be defined functionally by performing deletion analyses followed by stable transfection of cells (Philipsen *et al.*, 1993, *EMBO J* 12, 1077-85) or transgenic studies (Lowrey *et al*, 1992, *Proc Natl Acad Sci U S A* 89, 1143-7; These approaches define the minimum extent of sequence required to retain the biological function associated with the HS under examination.

It is observable that many hypersensitive sites occur within broader domains of increased DNase sensitivity and therefore appear to be components of higher-order chromatin structures. It is further observable that, based on published data, such sites

appear to harbor increased biological significance and are perhaps the most important functionally. Several investigators have observed that the regions flanking the hypersensitive foci of active elements exhibit an increased level of sensitivity to nuclease digestion compared with the increased general sensitivity of an active locus. This phenomenon has been referred to as 'intermediate sensitivity' (Kunnath and Locker, 1985, *Nucleic Acids Res.* 13; 115-29).

For more than two decades, the standard approach for measurement of chromatin accessibility has been nuclease hypersensitivity assays. In a conventional DNase hypersensitivity assay, intact nuclei are isolated from a cell type of interest and gently permeabilized. The nuclei are aliquoted and treated with with a series of increasing intensities of DNaseI (typically with increasing concentrations of the nuclease at fixed incubation time or alternatively with a fixed DNaseI concertration with increasing incubation times). The products are then deproeinated. Following DNA extraction and purification, samples from each aliquote are digested with a restriction enzyme, run over an agarose gel, and transferred to a membrane. To detect hypersensitive sites that are located within a particular restriction fragment, a probe is selected that is proximal to either the 5′ or 3′ end of the restriction fragment. Fragments are often probed from both ends to visualize cutting over both strands. Hybridization of a radiolabeled probe with the membrane highlights the parental band and sites that increase in intensity with increasing DNase concentration.

In spite of its extensively documented utility for localization of regulatory sequences, numerous technical barriers have prevented the broader application of conventional hypersensitivity assays to systematic detection of *cis*-active sequences on a genomic scale. The protocol (a) is extremely labor intensive; (b) is dependent on the presence of suitably-positioned restriction sites; (c) is further dependent on the availablility of a suitable ~500+bp sequence juxtaposed to a restriction site that can function as a specific probe (i.e., does not contain any repetitive sequences); (d) is highly consumptive of tissue resources, and therefore quite vulnerable to tissue preparation-to-preparation variability; (e) it suffers from numerous technical sources of variability including gel composition and running conditions, success of membrane transfer, success of probe labeling, hybridization conditions, wash conditions, and exposure conditions; and (f) it does not provide quantitative data. In practice, localization of the precise sequences which are hypersensitive is a difficult and laborious process requiring a series of restriction digests and probes positioned

immediately proximal to the site itself. Typically, probing from both sides of the site is desirable, and this process is necessary when more than one site is present on a given restriction fragment owing to a 'shadowing' effect by probe-proximal sites of those positioned more distally to the probe.

Significance of cis-regulatory sequences for studies of common diseases and environmental exposures

Inter-individual Variation in Gene Expression

Inter-individual variation in gene expression has been recognized for a number of human genes and is expected to underlie numerous quantitative phenotypes. For example, genes involved in xenobiotic metabolism and that of certain pharmaceutical agents (e.g., Cyp3A4, Cyp2, Thymidylate synthase, Nat1) are classical examples of enzymes that exhibit wide (up to 40- or even >100-fold) inter-individual variation in activity, much of which is attributable to transcriptional variation.

Several surveys have now documented the fact that a large proportion (at least 25%) of human genes are subject to such heritable variation in expression (Cheung *et al* 2002. *Nature Genet.* 32; 522-525, Schadt *et al.*, 2003. *Nature* 422; 297-302, Cheung *et al* .,2003. *Nature Genet.* 33, 422-425.). Comparable studies have also been performed in model organisms including the mouse (Cowles *et al.*, 2002. *Nature Genet.* 32; 432-437), *Fundulus* (Oleksiak *et al.*, 2002. *Nature Genet.*32; 261-266), and even yeast (Brem *et a.,l* 2002. *Science* 296; 752-755; Yvert *et al.*, 2003. *Nature Genet.* 35; 57-64.). Although elegantly executed, all of the aforementioned studies were capable of detecting only relatively large (>2-fold) changes in expression. Considerable data have emerged, however, to indicate that *in vivo,* even small differences in allelic expression can have dramatic phenotypic consequences. For example, a modest (<25%) decrease in total APC expression can result in a nearly 24-fold increase in risk of development of adenomatous polyposis coli and malignant lesions (Yan *et al.*, 2002. *Nature Genet.* 30; 25-26.). In the case of genes that exhibit a 'threshold' effect in activity (such as do many enzymes and receptors), the effect may be more pronounced. For example, even a 10% differences in the amount of CFTR transcript can dramatically attenuate the cystic fibrosis phenotype (Rave-Harel *et al.*, 1997. *Am. J. Hum. Genet.* 60; 87-94.; Ramalho *et al.*, 2002. *Am. J. Respir. Cell. Mol. Biol.* 27; 619-627).

Importance of cis-regulatory sequences for quantitative phenotypes

Common diseases are characterized by polygenic inheritance and by quantitative (i.e., continuous) variation in specific phenotypic traits. A major biological mechanism contributing to quantitative phenotypic variation is heritable variation in the regulation of gene expression. In humans, such variation is expected to reside principally within *cis*-regulatory sequences (Rockman and Wray 2002. *Mol. Biol. Evol.* 19; 1991-2004.). Since individual *trans*-regulatory transcriptional factors typically interact with a wide network of genes, variation affecting these proteins would be expected to have pleiotropic effects and comparatively dramatic phenotypes, and are therefore anticipated to be quite rare. An example of this phenomenon may be found in inherited defects in transcriptional factors which give rise to marked early-onset Type 2 diabetes (MODY) phenotypes (Lehto *et al.*, 1999. *Diabetes* 48; 423-425, Chang *et al.*, 1997. *Eur. J. Biochem.* 247; 148-159).

Since transcriptional factors require interaction with *cis*-regulatory sites in order for their effects to be manifest, defects in the genomic target sites of these factors may produce similar (though quantitatively more subtle) physiological consequences. However, the impact of *cis*-regulatory variations should directly impact only their cognate gene(s). *Cis*-regulatory variation could manifest functionally in a variety of ways by impacting (a) the magnitude of gene expression; (b) regulation of tissue-specificity; (c) control over timing of expression during development and differentiation; (d) response to environmental stimuli (such as pharmacologic agents); or (e) some combination thereof. Given the overall prevalence of human genetic variation, lesions in one or more of the cognate *cis*-regulatory sites should be comparatively common. When the multiple regulatory factors that interact with each regulatory sequence of each gene are considered, such *cis*-variation would provide the ideal substrate for a complex, semi-quantitatively varying phenotype.

There presently exist hundreds of reports in the literature of associations between genetic variation in known or suspected regulatory regions and phenotypic manifestations or disease risk (see extensive tabulations in Rockman and Wray 2002. *Mol. Biol. Evol.* 19; 1991-2004.; Haukim *et al.*, 2002. *Genes Immun.* 3; 313-330). Because the region immediately upstream of the transcriptional start site of human

genes often (though not universally) demarcates the proximal promoter region, it is not surprising that the vast majority of efforts to locate polymorphisms that impact transcriptional regulation have focused on this region. While it is tempting to conclude that any polymorphism within the upstream region of genes is regulatory in nature, this overlooks the fact that the specific sequences which are active *in vivo* – i.e., those to which transcriptional factors are complexed – are in fact highly compartmentalized into discrete domains of remodeled chromatin (Felsenfeld 1996. *Cell* 86; 13-19; Struhl 2001. *Science* 293; 1054-1055.). It is thus presently the case that many reports of regulatory polymorphism in the literature likely represent cases that would more correctly be classified simply as 'non-coding polymorphism of undetermined significance'. The availability of a molecular method capable of localizing actual cis-regulatory sequences would therefore have a major impact on studies of genetic variation.

Even in cases where functional documentation has been undertaken, the focus on the proximal upstream region has resulted in a significant ascertainment bias, which is reflected in the fact that nearly 80% of all documented regulatory polymorphisms described are found within the first 600bp upstream of transcription start sites (Rockman and Wray, 2002. *Mol. Biol. Evol.* 19; 1991-2004).

Quantitative variation in serum lipids. A clear illustration of the effect of regulatory polymorphism in modulating quantitative phenotypes is provided by serum lipids. An extensive literature has now emerged relating dyslipidemias with regulatory polymorphism in major apolipoprotein and lipolytic genes including *ApoA1* (Smith *et al.,* 1992. *J. Clin. Invest.* 89; 1796-1800; Barre *et al.,* 1994. *J. Lipid Res.* 35; 1292-1296; Juo *et al.*, 1999. *Am. J. Med. Genet.* 82; 235-241), *ApoC3* (Dammerman *et al.,* 1993. *Proc. Natl. Acad. Sci. USA* 90; 4562-4566; Hegele *et al.*, 1997. *Arterioscler. Thromb. Vasc. Biol.* 17; 2753-2758), *ApoB* (Van Hooft *et al.*, 1999. *J. Lipid Res.* 40; 1686-1694), *ApoE* (Nickerson *et al.*, 2000. *Genome Res.* 10; 1532-1545), *ApoC1* (Xu *et al.*, 1999. *J. Lipid Res.* 40; 50-58), hepatic lipase (Guerra *et al* ., 1997. *Proc. Natl. Acad. Sci. USA* 94; 4532-4537; Deeb and Peng 2000. *J. Lipid Res.* 41; 155-158; Zambon *et al.,* 2003. *Curr. Opin. Lipidol.* 14; 179-189; Murtomaki *et al.*, 1997. *Arterioscler. Thromb. Vasc. Biol.*17; 1879-1884), lipoprotein lipase (Hall *et al.*, 1997. *Arterioscler. Thromb. Vasc. Biol.*17; 1969-1976; Talmud *et al.,* 1998. *Biochem. Biophys. Res. Commun.* 252; 661-668;), hormone-sensitive lipase (Pihilajamaki *et a.,l* 2001. *Eur. J. Clin. Invest.* 31; 302-308; Talmud *et al.*, 1998. *J.*

Lipid. Res. 39; 1189-1196), and cholesterol esterase transfer protein (Dachet *et al.*, 2000. *Arterioscler. Thromb. Vasc. Biol.* 20; 507-515). Many of these functional polymorphisms had been further shown to influence atherosclerosis (Ye *et al.*, 1996. *J. Biol. Chem.* 271; 13055-13060; Jansen *et al.*, 1997. *Arterioscler. Thromb. Vasc. Biol.* 17; 2837-2842; Corbex *et al.*, 2000. *Nature Genet.* 32; 432-437), myocardial infarction (Lambert *et al.*, 2000. *Hum. Mol. Genet.* 9; 57-61; Ericksson *et al.*, 1995. *Proc. Natl. Acad. Sci. USA* 92; 1851-1855), and stroke (Ito *et al.*, 2000. *Stroke* 31; 2661-2664; Nakayama *et al.*, 2000. *Am. J. Hypertens.* 13; 1263-1267).

Regulatory polymorphism in common diseases with known or suspected environmental components

Compelling evidence now exists for the involvement of regulatory polymorphism in diverse diseases for which a major environmental component exists. Relevant examples include:

Pulmonary diseases. Regulatory polymorphism has recently emerged as a centerpiece of studies of the genetic determinants of airway reactivity, and has been described in several genes associated with asthma (In *et al.*, 1997. *J. Clin. Invest.* 99; 1130-1137; Silverman *et al.*, 1998. *Am J. Respir. Cell Mol. Biol.* 19; 316-323; Scott *et al.*, 1999. *Br. J. Pharamacol.* 126; 841-844; Drazen *et al.*, 1999. *Nature Genet.* 22; 168-170; Sanak *et al..*, 2000. *Am .J. Respir. Cell Mol. Biol.* 23; 290-296; Drysdale *et al.*, 2000. *Proc. Natl. Acad. Sci. USA* 97; 168-170), chronic respiratory disease (Morgan *et al.*, 1993. *Hum. Mol. Genet.* 2; 253-257) including COPD (Keatings *et al.*, 2000. *Chest* 118; 971-975) and environmental susceptibility to emphysema (Yamada *et al.*, 2000. *Am J. Hum. Genet.* 66; 187-195).

Allergic and autoimmune diseases. Functional non-coding polymorphisms have also been implicated in allergic (Nickel *et al.*, 2000. *J. Immunol.* 164; 1612-1616) and autoimmune diseases including juveline rheumatoid arthritis (Crawley *et al.*, 1999. *Arthritis Rheum.* 42; 1101-1108; Fishman *et al.*, 1998. *J. Clin. Invest.* 102; 1369-1376), SLE (Stevens *et al.*, 2001. *Arthritis Rheum.* 44; 2358-2366), myasthenia gravis (Kaluza *et al.*, 2000. *J. Invest. Dermatol.* 114; 1180-1183), systemic sclerosis (Hata *et al.*, 2000. *Biochem. Biophys. Res. Commun.* 272; 36-40), and Type I diabetes (Kennedy *et al.*, 1995. *Nature Genet.* 9; 293-298; Lew *et al.*, 2000. *Proc. Natl. Acad. Sci. USA* 97; 12508-12512; Pugilese *et al.*, 1997. *Nature Genet.* 15; 293-297).

Cancer. Regulatory polymorphisms in a variety of genes had been associated with cancers of the ovary (Phelan *et al.,* 1996. *Nature Genet.* 12; 309-311), aerodigestive tract (Cascorbi *et al.,* 2000. *Cancer Res.* 60; 644-649), lung (Zhu *et al .,* 2001. *Cancer Res.* 61; 7825-7829), endometrium (Nishioka *et al.*, 2000. 91; 612-615), prostate (Rebbeck *et al.*, 2000. *J. Natl. Cancer Inst.* 92; 76; Rebbeck *et al.,* 1998. *J. Natl. Cancer Inst.* 90; 1225-1229), and skin (Foster *et al.*, 2000. *Blood* 96; 2562-2567; Ye *et al.*, 2001. *Cancer Res.* 61; 1296-1298).

Common birth defects. At least one report has specifically connected regulatory polymorphism of PDGF-alpha with neural tube defects during gestation (Joosten *et al.,* 2001. *Nature Genet.* 27; 215-217).

Functional polymorphism in sequences mediating specific physiological responses

Regulatory factor recognition motifs within *cis*-regulatory elements can be said to comprise the components of 'nodes' in transcriptional regulatory networks. Mutations disrupting or otherwise modifying specific factor motifs may thus shed light on the physiological connections of multi-gene pathways. Regulatory polymorphism has been described in *cis*-regulatory sequences which are known to respond to specific physiological stimuli including insulin (Groenendijk *et al.*, 1999. *J. Lipid Res.* 40; 1036-1044; Waterworth *et al.*, 2000. *J. Lipid Res.* 41; 1103-1109), low-density lipoproteins (Eriksson *et al.,* 1998. *Arterioscler. Thromb. Vasc. Biol.* 18; 20-26), sterols (Yang *et al.*, 1998. *J. Lipid Res.* 39; 2054-2064), retinoic acid (Piedrafita *et al.*, 1996. *J. Biol. Chem.* 271; 14412-14420), and estrogen (Morgan *et al.,* 2000. *J. Hypertens.* 18; 553-557). Mutations in specific drug responsive elements (e.g., nifedipine) have also been described (Walker *et al.,* 1998. *Hum. Mutat.* 12; 289).

Gene induction is a well-described response to a variety of external stimuli, classically xenobiotics. Metabolism of diverse pharmaceuticals is also heavily influenced by inter-individual variation in expression of metabolizing genes. Among enzymes which are known to be impacted by regulatory polymorphism are acetylcholinesterase (Shapira *et al.*, 2000. *Hum. Mol. Genet.* 9; 1273-1281), glutathione-S-transferase (Coles *et al.*, 2001. *Pharmacogenetics* 11; 663-669), monoamine oxidase (Denney *et al.*, 1999. *Hum. Genet.* 105; 542-551; Sabol *et al.*, 1998. *Hum. Genet.* 103; 273-279), thymidylate synthase (Mandola *et al.*, 2003. *Cancer Res.* 63; 2898-2904), ornithine decarboxylase (Guo *et al.*, 2000. *Cancer Res.*

60; 6314-6317), and tyrosine hydroxylase (Albanese *et a.,l* 2001. *Hum. Mol. Genet.* 10; 1785-1792; Meloni *et al.*, 1998. *Hum. Mol. Genet.* 7; 423-428). Regulatory polymorphisms of several genes involved in alcohol metabolism have also been described (Chou *et al.*, 1999. *Alcohol Clin. Exp. Res.* 23; 963-968; Edenberg *et al.*, 1999. *Pharmacogenetics* 9; 25-30) and at least one has been linked with clinical alcoholism (Harada *et al.*, 1999. *Alcohol Clin. Exp. Res.* 23; 958-962).

Regulatory polymorphism also appears to be prevalent within p450 enzymes including *CYP1A2* (Aitchison *et al.*, 2000. *Pharamacogenetics* 10; 695-704), *CYP2E1* (Hayashi *et al.*, 1991. *J. Biochem.* 110; 559-565; Watanabe *et al.*, 1994. *J. Biochem.* 116; 321-326; Hildesheim *et al.*, 1995. *Cancer Epidemol. Biomarkers Prev.* 4; 607-610; Fairbrother *et al.*, 1998. *Pharmacogenetics* 8; 543-552; Marchand *et al.*, 1999. *Cancer Epidemol. Biomarkers Prev.* 8; 495-500; Chabra *et al.*, 1999. *Carcinogenesis* 20, 1031-1034), *CYP2A6* (Pitarque *et al.*, 2001. *Biochem. Biophys. Res. Commun.* 284; 455-460), and *CYP3A4* (Rebbeck *et al.*, 1998. *J. Natl. Cancer. Inst.* 90; 1225-1229; Amirimani *et al.*, 1999. *J. Natl. Cancer. Inst.* 91; 1588-1590; Rebbeck 2000. *J. Natl. Cancer. Inst.* 92; 76).

The aforementioned examples provide powerful evidence of the existence and physiological relevance of regulatory polymorphism affecting a wide spectrum of human genes.

While promoter sequences are clearly necessary for expression, a recurring theme in the study of human gene regulation is that promoters alone are typically not sufficient either for high-level expression, nor for tissue-specific expression (or both). The Cyp3A genes catalyze the metabolism of structurally diverse endobiotics, drugs, and protoxic and procarcinogenic molecules and provide a relevant example. These genes exhibit substantial (>30-fold) interindividual variability in expression which is linked *in cis*. However, comprehensive sequencing of their promoter regions has thus far failed to disclose the responsible molecular lesions (Kuehl et al 2001). The distal regulatory sequences of *Cyp3A* genes have not been delineated. ***This example provides clear rationale for the necessity of searching for polymorphism in distal regulatory sequences.***

Because of the difficulty in locating distal regulatory sequences using conventional methods, however non-promoter regulatory variants have not been amenable to systematic study. Nonetheless, several cases of non-promoter regulatory

polymorphism have come to light, often with clear clinical correlates. Examples include alpha1 immunoglobulin (Denizot et al 2001), ornithine decarboxylase (Martinez *et al.*, 2003. *Proc. Natl. Acad. Sci. USA* 100; 7859-7864), apolipoprotein(a) (Wade *et al.*, 1991. *Atherosclerosis* 91; 63-72; Wade *et al.*, 1994. *J. Biol. Chem.* 269; 19757-19767; Wade *et al.*, 1997. *J. Biol. Chem.* 272; 30387-30399; Puckey and Knight 2003. *Atherosclerosis* 166; 119-127), the Calpain10 gene implicated in Type2 diabetes (Horikawa *et al.*, 2000. *Nature Genet.* 26; 163-175; Cox 2001. *Hum. Mol. Genet.* 20; 2301-2305), the Renin gene enhancer (Fuchs *et al.*, 2002. *J. Hypertens.* 20; 2391-2398); and an intronic enhancer of *PDCD1*, associated with development of systemic lupus erythematosus (Prokunina *et al.*, 2002. *Nature Genet.* 32; 666-669). A functional lesion within a regulatory sequence located >17kb distant to the acetylcholinesterase gene has been identified characterized *in vivo* (Shapira *et al.*, 2000. *Hum. Mol. Genet.* 9; 1273-1281). ***The example of acetylcholinesterase provides further proof-of-principle for the existence of functional polymorphism in distant regulatory sequences that have pronounced and heritable phenotypic manifestations.***

Regulatory polymorphisms may also interact with protein coding lesions to potentiate or ameliorate their phenotypic consequences. Examples of this phenomenon are found in *CFTR* (Romey *et al.*, 1999. *J. Med. Genet.* 36; 263-264; Romey *et al.*, 2000. *J. Biol. Chem.* 275; 3561-3567; Romey *et al.*, 1999. *Hum. Genet.* 105; 145-150) and in LTA, where co-occurrence of a functional intronic enhancer polymorphism and a non-synonymous coding variant substantially increase the risk of myocardial infarction in homozygotes (Ozaki *et al.*, 2002. *Nature Genet.* 32; 650-654).

These examples and others highlight the value of the approach we propose to employ in this study, namely, targeted interrogation of candidate* cis*-regulatory sequences to discover functional regulatory alleles that may modulate important clinical traits and disease phenotypes. The fact that examples of extra-promoter regulatory polymorphism such as the above have come to light in spite of the limited database of known distal regulatory sequences highlights the promise of systematic, large-scale mining of such elements over a gene set of broad physiological relevance.

Comparatively 'deep' surveys of genetic variation are a logical approach to regions of the genome in which polymorphisms would be expected to alter gene function or expression, and thereby contribute to phenotypic variation. Polymorphisms with functional consequences are expected to have lower allele frequencies and, in fact, the majority of coding region SNPs (cSNPs) that change an amino acid have allele frequencies below 5% (Cargill *et al.*, 1999. *Nature Genet.* 22; 231-238; Halushka *et al.*, 1999. *Nature Genet.* 22; 239-247). Target population sizes sufficient for comprehensive identification of alleles with frequencies of 1-5% are therefore most desirable and have motivated the sample sizes used in this proposal.

Cis-regulatory regions are of the greatest scientific and clinical interest though they are extremely difficult to delineate and study using conventional approaches. Identification of regulatory regions is expected to be of central importance to our understanding of common diseases, quantitative traits, and environmental exposures.

Computational approaches to the study of *cis*-regulatory sequences

Overview. The search, via computational methods, for cis-regulatory elements in genomic DNA has been pursued using three different classes of techniques: motif discovery algorithms, algorithms for recognizing cis-regulatory modules, and non-motif-based algorithms. The problem is particularly challenging in the human genome, owing not only to its size and sequence diversity, but mainly to the fact that human gene regulation is characterized by coordinate action of multiple *cis*-regulatory elements over distances of many kilobases.

Algorithms for *de novo* discovery of TFBS motifs

The first class of algorithms performs *de novo* discovery of transcription factor binding site (TFBS) motifs in relatively small sets of DNA sequences. This class includes algorithms such as the Gibbs sampler (Lawrence et al., 1993. Science, 262(5131):208-214), MEME (Bailey and Elkan, 1994. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pages 28-36) and Consensus (Hertz and Stormo, 1999. Bioinformatics, 15(7):563-577).Recent research in this area focuses on building richer motif models (Xing et al., 2003. Advances in Neural Information Processing Systems, Cambridge, MA, 2003. MIT

Press), on developing provably optimal algorithms (Eskin et al., 2003. Proceedings of the Pacific Symposium on Biocomputing, pages 29-40, New Jersey, 2003. World Scientific), on finding pairs of co-occurring binding sites (Eskin and Pevzner, 2002. Bioinformatics, 18: S354-S363, van Helden et al., 2000. Nucleic Acids Research, 28(8):1808-1818), and on searching simultaneously with sequence information and other types of data(Loots et al., 2002. *Genome Res.* 12, 832-9, Blanchette and Tompa, 2002. Genome Research, 12(5):739-748, McCue et al., 2001. Nucleic Acids Research, 29(3): 774-782. , Bussemaker et al., 2001. Nature Genetics, 27:167-171, Holmes and Bruno, 2000. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, pages 202-210). However, because these algorithms are appropriate only for relatively small data sets, they all require prior knowledge of the approximate locations of a collection of similar TFBS's.

Algorithms for discovery of cis-regulatory modules

Algorithms in the second class, in contrast, operate on much larger sequence databases; however, these algorithms generally assume that the statistical properties of a small collection of transcription factor binding sites are known *a priori.* Here, the problem is to locate statistically significant clusters of these binding sites, called regulatory modules, in genomic DNA. Three groups of algorithms for recognizing regulatory modules have been proposed. Algorithms in the first group use a sliding window approach, scoring each subsequence that appears in the window with respect to a given collection of motifs (Prestridge, 1995. Journal of Molecular Biology, 249:923-932, Kondrakhin et al., 1995. Computer Applications in the Biosciences, 11:477-488, Frech et al.,1997. Journal of Molecular Biology, 270: 674-687, Berman et al., 2002. *Proc Natl Acad Sci USA*, 99:757-762, Markstein et al., 2002. Proc Natl Acad Sci U S A. 99:763-8, Levy and Hannenhalli, 2002. Mammalian Genome, 13:510-514, Johansson et al., 2003. Bioinformatics, 19(Suppl. 1):i169-i176, Sharan et al., 2003. Bioinformatics, 19(Suppl. 1):i292-i301). The sliding window approach has intuitive appeal, and has yielded good results in analyses of motif clusters in Drosophila (Berman et al., 2002. Proceedings of the National Academy of Sciences of the United States of America, 99:757-762, Markstein et al., 2002. Proc Natl Acad Sci U S A. 99:763-8). The second group of search algorithms uses a probabilistic modeling framework called hidden Markov models (HMMs) (Frith et al., 2001.

Bioinformatics, 17(10):878-889, 2002, Bailey and Noble, 2003. Bioinformatics, 19(Suppl. 2):ii16-ii25). The HMM approach is more theoretically rigorous and offers more accurate statistics than the relatively ad hoc sliding window approach. However, both the sliding window and the HMM approaches to the regulatory module search problem are generative: both rely upon a model (implicit or explicit)of a regulatory module. The third group of algorithms uses a discriminative technique. These methods model the difference between the regulatory module and non-regulatory sequence. Logistic regression analysis (LRA)is a discriminative technique based upon a sliding window, which has been used successfully to build predictors for muscle-specific (Wasserman and Fickett, 1998. Journal of Molecular Biology, 278:167-181) and liver-specific (Krivan and Wasserman, 2001. Genome Research, 11:1559-1566) regulatory modules. The Fisher kernel support vector machine (SVM) method (Pavlidis et al., 2001. Proceedings of the Pacific Symposium on Biocomputing, pages 151-163) uses a discriminative algorithm based upon a hidden Markov model. In the presence of a small amount of data, discriminative techniques typically achieve better performance than similar, generative techniques.

Non-motif-based methods

The third class of algorithms for identifying cis-regulatory elements is the most general, requiring as input only a database of genomic DNA and producing as output, for example, the predicted locations of promoter regions or CpG islands. Many techniques in this class are non-motif based, capitalizing instead on compositional statistics (see Zhang (2002) Nature Reviews Genetics, 3:698-710, for a review). Some methods augment these statistics using libraries of known TFBS's (Crowley et al., 1997. Journal of Molecular Biology, 268:8-14) or libraries of words extracted in an unsupervised fashion from sequence databases (Scherf et al., 2000. Journal of Molecular Biology, 297:599-606). While most promoter recognition techniques are generative, at least one discriminative method has been described (Davuluri et al., 2001. Nature Genetics, 29(4):412-417).

Data fusion

Increasingly, the analysis of regulatory elements in DNA faces problems related to *data fusion*, i.e., drawing inferences from a collection of heterogeneous

data. For any of the search problems described above, a solution that operates only on the given DNA sequences suffers from a loss of power relative to a competing method that capitalizes on various types of auxiliary data. The simplest approach to data fusion is to treat each type of data independently. For example, co-expression of genes in microarray experiments may be used to select a collection of upstream regions for analysis by a motif discovery algorithm (Chu et al., 1998. Science, 282:699-705). Similarly, conservation of human DNA with respect to the mouse genome may be used to reduce the size of a database to be scanned. More powerful techniques learn simultaneously from two or more types of data, e.g., from DNA sequence and microarray data (Bussemaker et al., 2001 Nature Genetics, 27:167-171), or from DNA from multiple species (Duret and Bucher, 1997. Current Opinions in Structural Biology, 7:399-405, Blanchette and Tompa, 2002. Genome Research, 12(5):739-748). Indeed, the problem of discovering motifs in the presence of multi-species sequence data is called phylogenetic footprinting (Tagle et al., 1988. . Journal of Molecular Biology, 203:439-455) and has recently seen success in an analysis of four yeast genomes (Kamvysselis et al., 2003. In Proceedings of the Seventh Annual International Conference on Computational Molecular Biology, pages 157-166; Kellis et al 2003. Nature 423:241-54).

In vivo molecular validation of computational predictions

To date, there have been few published efforts to perform *in vivo* validation of computational predictions, owing mainly to the painstaking nature and cost of conventional molecular methodologies. All have been performed in lower-complexity genomes than the human, principally *Drosophila* (see references above) and *C. elegans* (Gaudet et al 2002. Science 295(5556):821-5), and generally under idealized situations such as a restricted developmental window when the action of specific morphogenic transcription factors predominates. Furthermore, all published studies have relied on motif-based approaches and it is observable that the findings forthcoming from the majority have pertained to homotypic regulatory elements (i.e., those which contain clusters of a binding sites for single transcriptional factor). Finally, the predicted sensitivity of the approaches is poor, since only a few dozen statistically-significant predictions were made even in genome-wide searches.

Significantly, in no case has any computational methodology undergone rigorous in vivo validation sufficient to establish (or reject) its predictive value.

Use of comparative genomic approaches to predict regulatory sequences

Comparative genomic analyses represent a conceptually attractive approach for identification of regulatory sequences (Ureta-Vidal et al. 2003. *Nat. Rev. Genet.* 4, 251-62). The central hypothesis of such studies is that functionally important sequences will exhibit selective pressures that propagate over evolutionary distances (Dermitzakis et al. 2002. *Nature* 420, 578-82). However, in reality the situation is complex. For example, while it is clear that certain regulatory elements have been highly conserved during vertebrate and particularly mammalian evolution (Elnitski et al. 2003. *Genome Res.* 13, 64-72), it is also evident that many such elements exhibit little or no selective conservation above local background (Flint et al. 2001. *Hum. Mol. Genet.* 10, 371-82).

Given that a surprisingly large proportion of the human genome appears to be under selection (Waterston et al 2002. Nature 420(6915):520-62), the task that we ask of a comparative genomics-based method is: *can functional elements in the human genome be reliably and specifically discriminated from background levels of conservation?* To date, there is little evidence that this can be accomplished in a manner that displays adequate sensitivity, specificity, and generalizes well across the genome. The number of studies evaluating elements identified purely on the basis of comparative genomics (predomintly mouse-human) approaches are very few and in no case has the comparative genomic hypothesis been rigorously examined. Furthermore, an interesting feature of several such studies is the fact that the elements which were reported to be identified on the basis of comparative genomics had in fact been reported previously to be DNaseI hypersensitive sites (Loots et al 2002. Genome Research, 12(5):832-839; Mohrs et al 2001; Gottgens et al 2000. Nat Biotechnol. 18(2):181-6.). For example, in one study of the interleukin cluster on chromosome 5 (Loots et al 2002 Genome Research, 12(5):832-839), 90 conserved non-coding sequences were identified, but the only one was selected for *in vivo* studies was in fact a previously described and studied DNaseI hypersensitive site (Takemoto et al 1998. Int Immunol. 10(12):1981-5).

The recent availability of comparative sequence information from a range of vertebrate and mammalian species has now made practical the description and evaluation of sequence elements conserved across multiple species (so-called multi-species-conserved elements or 'MCSs' (Thomas et al 2003. Nature 424(6950):788-93)). However, although this information imparts some specificity, it does not seem to impact the sensitivity as evidenced by poor performance in identifying previously-characterized regulatory elements. For example, only a small fraction of the numerous DNaseI hypersensitive sites identified within and flanking the *CFTR* gene (Nuthall et al 1999a. Biochem J. 1999 341 (Pt 3):601-11; Nuthall et al 1999b. Eur J Biochem. 1999 266(2):431-43; Smith et al 2000. Genomics 64(1):90-6) were found to coincide with MCSs, in spite of the fact that hundreds of MCSs were identified in this region.

The availability of a generic high-throughput, in vivo functional method to identify candidate regulatory sequences would obviate the need to rely on comparative analyses as a primary discovery vehicle. Rather, their value could be realized mainly by further illumination of functionally-derived information. Such a functional method is described below and will be applied in the proposal.

A method for quantitatively determining DNA sensitivity to DNA modifying agents has been disclosed in PCT publication WO 02/097135 and McArthur *et al.*, 2001, *J Mol Biol* 313, 27-34.

Discussion or citation of a reference herein shall not be construed as an admission that such reference is prior art to the present invention.

3. SUMMARY OF THE INVENTION

The invention provides a method for profiling chromatin sensitivity of a genomic region of cells of a cell type to digestion by a DNA modifying agent. The method comprises determining a chromatin sensitivity profile, said chromatin sensitivity profile comprising a plurality of replicate measurements of each of a plurality of different genomic sequences in said genomic region, wherein each of said plurality of replicate measurements is a ratio of (i) copy numbers of an amplicon comprising said genomic sequence measured by real-time quantitative PCR (qPCR) with chromatin of said cell type that has been treated with said DNA modifying agent and (ii) copy numbers of said amplicon measured by real-time qPCR with chromatin of said cell type that has not been treated with said DNA modifying agent. In a

preferred embodiment, the plurality of different genomic sequences comprises successively overlapping sequences tiled across one or more portions of said genomic region. In another preferred embodiment, the plurality of different genomic sequences comprises successively overlapping sequences tiled across said genomic region. Preferably, the plurality of different genomic sequences has a length in the range of about 75 to about 300 bases. In one preferred embodiment, the mean length of said plurality of different genomic sequences is about 250 bases. Preferably, the plurality of duplicate measurements for each genomic sequence in the chromatin sensitivity profile consists of at least 3, at least 6, or at least 9 duplicate measurements.

In another embodiment, the invention provides a method for profiling chromatin sensitivity of a genomic region of cells of a cell type to digestion by a DNA modifying agent, comprising (a) treating chromatin of cells of said cell type with said DNA modifying agent such that digestion of DNA occurs and retrieving DNA molecules; (b) amplifying a plurality of different genomic sequences in said genomic region by real-time quantitative PCR using at least a portion of said retrived DNA molecules and determining copy numbers of amplification product of each said genomic sequence; (c) amplifying said plurality of different genomic sequences in said genomic region by real-time quantitative PCR using DNA molecules obtained from chromatin of cells of said cell type that is not treated by said DNA modifying agent and determining copy numbers of amplification product of each said genomic sequence; (d) determining a ratio of said copy numbers measured in step (b) and copy numbers measured in said step (c); (e) repeating said steps (b) - (d) a plurality of times to generate a plurality of ratios, thereby generating a plurality of replicate measurements for each of said genomic sequences; and (d) determining a chromatin sensitivity profile of said genomic region, said chromatin sensitivity profile comprising said plurality of replicate measurements.

In a preferred embodiment, copy numbers are corrected for difference in amplification efficiency. In another preferred embodiment, the DNA modifying agent is DNase I. The plurality of duplicated measurements can be measured by independent real-time qPCR experiments. The plurality of duplicated measurements can also measured by independent real-time qPCR experiments using different treated chromatin samples.

In certain illustrative embodiments, the genetic locus being profiled will contain at least one coding region for at least one expressed gene, for example a known gene having a putative or assigned association with a disease state or other abnormal cellular condition. In one preferred embodiment, for example, the methods of the invention are employed to generate RS profiles for genes associated with cancer. Such genes can include essentially any gene known or believe to be associated with cancer, including, for example, genes such as p53, Rb, INK4A/p16, CTNNB1, H-Ras, Fos, MDM2, INK4, ARF1, PTEN, Jun, WNT3A/14, NFkB, TERT, BRCA1, BRCA2, WAF1/p21, CDK4, TGF-beta1, RAR, E2F, VHL, MLH1, SMAD4, SMAD2, SMAD3, K-Ras, EGFR, WT1, Myc, Raf, ABL, HER2.

The genetic loci profiled according to the present invention may comprise essentially any size of genetic material provided the locus is of sufficient length to allow for the identification of regulatory sequences within said locus. Typically, the genetic locus to be profiled according to the methods of the present invention will comprise greater than about 1 kb of DNA, greater than about 10 kb of DNA, greater than about 25 kb of DNA, greater than about 50 kb of DNA or greater than about 100 kb of DNA. Thus, in most instances, the genetic locus profiled according to the present invention will comprise about 1 to 100 kb of DNA, about 25 to 75 kb of DNA, or about 50 to 100 kb of DNA.

The step of identifying regulatory sequences associated with the genetic locus being profiled can be carried out according to essentially any methods known and available in the art. In a particularly illustrative embodiment, the step of identifying regulatory sequences associated with the genetic locus being profiled is performed by a plurality of polymerase chain reactions using primers that amplify products that overlap and span substantially the entirety of the genetic locus of interest. As described further herein, this allows for a rapid and high throughput means to identify and characterize regulatory sequences present within the genetic locus.

In one embodiment, the primers used in the plurality of PCR reactions are designed so as to amplify products comprising DNA sequences having lengths between about 100 and 1000 base pairs, between about 100 and 500 base pairs. In certain embodiments, it is preferred that the amplified products have length between about 200 and 300 base pairs.

Agents that induces modifications in DNA at hypersensitivity sites are known and available in the art, illustrative examples of which may be selected from the group

consisting of radiation, such as light radiation, a chemical agent, such as a clastogen, an enzyme, and combinations thereof.

The enzymes employed in this regard may selected from essentially any enzyme capable of modifying DNA at hypersensitivity sites, and most typically will be selected the group consisting of specific endonucleases, non-specific endonucleases, topoisomerases, methylases, histone RStylases, histone deRStylases, and combinations thereof. Certain illustrative specific endonucleases comprise one or more four-base restriction endonucleases, one or more six-base restriction endonucleases, or combinations thereof. Certain illustrative four-base restriction endonucleases may selected from the group consisting of *Sau*3a, *Styl*, *Nla* III, *Hsp* 92, and combinations thereof. Certain illustrative six-base endonucleases may be selected from the group consisting of *Eco*RI, *Hind*III, and combinations thereof. In a particularly illustrative embodiment, the enzyme is a non-specific endonuclease, preferably DNase I.

In another embodiment, the method of the invention further comprises determining a baseline chromatin sensitivity profile by a method comprising (a) smoothing the data in said chromatin sensitivity profile to obtain a baseline curve; and (b) determining the error bounds for said baseline curve, wherein said baseline curve and said error bounds constitute said baseline chromatin profile. Preferably, the smoothing is carried out using LOWESS. In one embodiment, the error bounds are determined by a method comprising (b1) mean centering said plurality of replicates for each genomic sequence in said chromatin sensitivity profile about said baseline curve to generate a mean-centered chromatin sensitivity profile, wherein said mean-centering is carried out by setting the mean of each said plurality of replicates to the value of the corresponding genomic sequence on said baseline curve; (b2) determining the median M of said mean-centered chromatin sensitivity profile; (b3) determining the Median Absolute Deviation MAD of said mean-centered chromatin sensitivity profile; (b4) discarding for each genomic sequence replicate measurement X if X satisfy equation

$$\frac{|X-M|}{MAD/0.6745} > 2.24, \text{ and}$$

(b5) defining the error bounds as the lower and upper confidence limits on the remaining data.

In another embodiment, the error bounds are determined by a method comprising (b1) generating a bootstrap chromatin sensitivity profile by randomly selecting one replicate measurement from said plurality of replicate measurements for each genomic sequence; (b2) mean centering said plurality of replicates for each genomic sequence in said bootstrap chromatin sensitivity profile about said baseline curve to generate a mean-centered chromatin sensitivity profile, wherein said mean-centering is carried out by setting the mean of each said plurality of replicates to the value of the corresponding genomic sequence on said baseline curve; (b3) determining the median M of said mean-centered chromatin sensitivity profile; (b4) determining the Median Absolute Deviation MAD of said mean-centered chromatin sensitivity profile; (b5) discarding for each genomic sequence replicate measurement X if X satisfy equation

$$\frac{|X-M|}{MAD/0.6745} > 2.24,$$

(b5) determining the maximum lower and minimum upper outliers on the remaining data; (b6) repeating said step (b1)-(b5) for a plurality of times; and (b7) calculating the upper and lower outlier cutoff values and Bca confidence intervals.

In still another embodiment, the method further comprises (c1) identifying one or more genomic sequences among said plurality of genomic sequences whose Y% trimmed means lie outside said error bounds; and (c2) determining a signal-to-noise ratio S/N of said identified genomic sequences according to equation

$$S/N_i = \frac{|HS_i - B_i|}{MAD_B(\sigma_c/\sigma_{HS})^2}$$

where S/N_i is the signal-to-noise ratio at site i, HS_i is the Y% trimmed mean of the corresponding HS cluster, B_i is the value of said baseline curve at said site i, MAD_B is the median average deviation of the centered baseline, σ_{HS} is the average variance of replicate measurements, and σ_c is the variance of the replicate measurements at said site i. In one embodiment, the Y% trimmed mean is 20% trimmed mean.

According to another aspect of the present invention, there are provided regulatory sequence profiles identified according to the method of the present invention.

According to another aspect of the present invention, there are provided nucleotide arrays comprising a plurality of regulatory sequence sequences identified

by the methods of the present invention, wherein the array is fixed to a slide, a chip, or a membrane filter, for example.

According to another aspect of the present invention, there are provided methods for ascertaining the effect of an agent or other environmental perturbation on an regulatory sequence profile of a genetic locus by obtaining a first regulatory sequence profile associated with the genetic locus, wherein the sample from which the regulatory sequences are identified is unexposed to the agent or perturbation; obtaining a second regulatory sequence profile associated with the genetic locus, wherein the sample from which the regulatory sequences are identified is exposed to the agent or perturbation; and comparing the first profile with the second profile to determine regulatory sequences that are effected by the agent perturbation.

In certain illustrative embodiments of this aspect of the invention, the perturbation occurs before obtaining the sample from a tissue, wherein the environmental perturbation is selected from the group consisting of an infection of the eukaryotic organism from a microorganism, loss in immune function of the eukaryotic organism, exposure of the tissue to high temperature, exposure of the tissue to low temperature, cancer of the tissue, cancer of another tissue in the eukaryotic organism, irradiation of the tissue, exposure of the tissue to a chemical or other pharmaceutical compound; and aging.

In certain other embodiment according to this aspect of the invention, the perturbation occurs after obtaining the sample from a tissue, wherein the perturbation is selected from the group consisting of exposure of the tissue to high temperature, exposure of the tissue to low temperature, irradiation of the tissue, exposure of the tissue to a chemical or other pharmaceutical compound, and aging.

According to yet another aspect of the present invention, there are provided methods for profiling differential regulatory sequence activation associated with a genetic locus, comprising first obtaining multiple regulatory sequences associated with the genetic locus from a first population and labeling them with a first label; obtaining multiple regulatory sequences associated with the genetic locus from a second population and labeling them with a second label; hybridizing the elements with a DNA microarray containing DNA species in separate locations that match putative or verified regulatory elements associated with the genetic locus; and determining the ratio of signals from the first and second labels within the array. In a particularly illustrative embodiment, one of the populations is an untreated control

and the other population is treated by contact with at least one agent, and the signal ratios obtained provide an indication of gene regulatory activity modulated by the agent.

The invention further provides methods of using regulatory sequences profiles for a variety of purposes related generally to gene regulation, cell characterization and identification of drugs and therapies.

In one embodiment, the invention provides a method of identifying a gene associated with a disease or disorder, comprising comparing an regulatory sequence profile of a cell with a disease or disorder to an regulatory sequence profile of a normal control cell, identifying an regulatory sequence with different activities in the two cells, and identifying a gene associated with the identified regulatory sequence. In one embodiment, the active chromatin profiles are associated with a known gene or a specific chromatin region. In one embodiment, the disease or disorder is a cancer. In certain embodiments, the comparison is performed using an array of regulatory sequence sequences. The array may include regulatory sequence sequences associated with a plurality of genes.

In a related embodiment, the invention includes a method of identifying an regulatory sequence of a gene, comprising preparing an regulatory sequence profile of a gene and identifying an regulatory sequence within the profile. The regulatory sequence profile is prepared according to the method of claim 1.

In yet another embodiment, the invention includes a method of identifying an allelic form of a gene, comprising comparing an regulatory sequence profile of one cell to an regulatory sequence profile of a second cell, wherein the regulatory sequence profiles are associated with the same gene and identifying an regulatory sequence displaying different activities in the two cells. The method may further comprise obtaining the sequence of at least one of the identified regulatory sequences.

Another embodiment provides a method of identifying a cell, comprising determining the regulatory sequence profile associated with a cell, comparing the regulatory sequence profile of the cell to an regulatory sequence profile associated with a known cell types; and identifying a cell type with the same or a substantially similar regulatory sequence profile as the cell, thereby identifying the cell type of the cell. In one embodiment, the comparison is performed using an array of polynucleotides comprising regulatory sequences.

Another embodiment of the invention provides a method of detecting a disease or disorder in a subject, comprising identifying an regulatory sequence profile associated with a disease or disorder; determining an regulatory sequence profile of a subject; and comparing the regulatory sequence profile of the subject to the regulatory sequence profile associated with the disease or disorder, wherein the same or a similar regulatory sequence profile indicates the presence of the disease or disorder, and wherein the regulatory sequence profiles are associated with the same genetic locus.

Another embodiment provided by the invention is a method of qualifying a patient for a clinical trial, comprising identifying an regulatory sequence profile of a patient, and comparing the regulatory sequence profile of the patient to an regulatory sequence profile identified in patients suitable for a clinical trial, wherein the regulatory sequence profiles are associated with the same genetic locus.

A related embodiment provides a method of selecting a therapy for a patient, comprising identifying an regulatory sequence profile of a patient, comparing the identified regulatory sequence profile to the regulatory sequence profile associated with a favorable outcome following a therapy; and selecting the therapy if the regulatory sequence profiles are the same or substantially similar.

Yet another related embodiment of the invention is a method of predicting the outcome of a disease or treatment protocol, comprising identifying an regulatory sequence profile of a patient, comparing the regulatory sequence profile identified in step (a) to the regulatory sequence profiles associated with one or more outcomes associated with a disease or treatment, and an regulatory sequence profiles associated with an outcome associated with a disease or treatment that is the same or substantially similar to the identified regulatory sequence profile.

A further embodiment of the invention is a method of screening a drug candidate, comprising identifying one or more regulatory sequence profiles associated with a cell with a disease or disorder, wherein the cell is not treated with a candidate drug, providing the candidate drug to a cell with the disease or disorder, identifying one or more regulatory sequence profiles associated with the cell provided with the candidate drug, and comparing the regulatory sequence profiles of steps (a) and (c) and thereby determining whether treatment with the candidate drug altered an regulatory sequence profile.

Another embodiment of the invention provides a method of identifying a drug useful in treating a disease or disorder, comprising identifying an regulatory sequence

profile associated with a disease or disorder, treating a cell with the disease or disorder with a candidate drug, identifying an regulatory sequence profile after treatment with the candidate drug, wherein the regulatory sequence profiles correspond to the same genetic locus; and comparing the regulatory sequence profiles to determine if treatment with the candidate drug affected the regulatory sequence profile.

In a related embodiment, the invention also provides a drug identified by a method of the invention.

Another embodiment of the invention is a method of manufacturing a drug, comprising identifying a drug that alters an regulatory sequence profile associated with a disease or disorder and manufacturing the identified drug.

The invention further provides, in other embodiment, a variety of computer readable medium and programs, which may be employed in identifying, characterizing and performing methods of the invention, for example.

In one embodiment, the invention provides a computer readable medium comprising an regulatory sequence profile associated with a genetic locus. In one embodiment, the genetic locus comprises an open reading frame. In one embodiment, the open reading frame encodes a gene associated with a disease or disorder. In certain embodiment, the disease or disorder is a cancer. In another embodiment, the gene is p53, Rb, INK4A/p16, CTNNB1, H-Ras, Fos, MDM2, INK4, ARF1, PTEN, Jun, WNT3A/14, NFkB, TERT, BRCA1, BRCA2, WAF1/p21, CDK4, TGF-beta1, RAR, E2F, VHL, MLH1, SMAD4, SMAD2, SMAD3, K-Ras, EGFR, WT1, Myc, Raf, ABL, or HER2. In certain embodiments, the active chromatin profile contains the genomic position and activity of one or more regulatory sequences. In another embodiment, the genetic locus comprises an open reading frame.

The invention further provides, in another related embodiment, a computer readable medium comprising a plurality of regulatory sequence profiles associated with a specific cell. In certain embodiments, the cell is a mammalian cell. In one embodiment, the cell is a diseased cell. In another embodiment, the regulatory sequence profiles include the genetic location and activities of at least one regulatory sequence.

In yet another embodiment, the invention includes a computer readable medium comprising a plurality of regulatory sequence profiles associated with different cells. In one embodiment, the regulatory sequence profiles are associated

with the same genetic locus. In one specific embodiment, the regulatory sequence profiles include regulatory sequence profiles associated with a plurality of genetic loci for each cell. In another specific embodiment, one or more cells is treated with an agent, which may be a drug candidate. In another embodiment, the cells are derived from different tissues. In one specific embodiment, one or more cells is a diseased cell.

Another embodiment of the invention is a computer readable medium comprising regulatory sequence profiles for at least two genetic loci, wherein each locus comprises an open reading frame and one or more regulatory sequences associated with that gene, and wherein the profile includes polynucleotide sequences which are sequences of open reading frames, sequences that hybridize to a an open reading frame under moderately stringent conditions, degenerate sequences of open reading frames, or sequences that hybridize to degenerate sequences of open reading frames. In one embodiment, the computer readable medium comprises the sequences for at least p53, Rb, INK4A/p16, CTNNB1, H-Ras, Fos, MDM2, INK4, ARF1, PTEN, Jun, WNT3A/14, NFkB, TERT, BRCA1, BRCA2, WAF1/p21, CDK4, TGF-beta1, RAR, E2F, VHL, MLH1, SMAD4, SMAD2, SMAD3, K-Ras, EGFR, WT1, Myc, Raf, ABL, or HER2. In a specific embodiment, at least one regulatory sequence is a promoter or enhancer of transcription for a gene.

Another embodiment of the invention provides a computer executable program for comparing regulatory sequence profiles of two or more cells, comprising inputting an regulatory sequence profile associated with a genetic locus in a first cell, inputting an regulatory sequence profile associated with the same genetic locus in a second cell, and outputting a comparison of the regulatory sequence profiles.

In yet another related embodiment, the invention provides a computer executable program for the identification of a cell, comprising inputting an regulatory sequence profile associated with one or more genetic loci in a cell, searching a data set comprising regulatory sequence profiles for the same genetic loci in one or more known cell types, and outputting a cell type with the same or a substantially similar regulatory sequence profile as the regulatory sequence profile.

Another embodiment of the invention includes a method of regulating gene expression, comprising identifying an regulatory sequence profile associated with a desired pattern of gene expression, preparing a nucleic acid vector comprising at least a plurality of regulatory sequences within the profile of step (a) operably linked to a

gene sequence, and introducing the vector into a cell. In a specific embodiment, the cell is stably introduced into the cell to obtain permanent heritable transmission of the regulatory sequences and operably linked gene sequence. In another specific embodiment, the gene encodes a regulatory protein. In another embodiment, the gene encodes a therapeutic molecule. In certain embodiments, the therapeutic molecule is a polypeptide or a polynucleotide, and in specific embodiments, the therapeutic molecule is selected from the group consisting of: ribozymes, antisense RNA, double-stranded RNA, small interfering RNA, and short hairpin RNA.

In another embodiment, the invention includes an regulatory sequence identified by a method of the invention.

In yet another embodiment, the invention includes an allelic variant identified by a method of the invention.

The invention further provides a computer executable program for profiling a genetic locus for active chromatin, comprising inputting data comprising regions of chromatin hypersensitivity sites derived from a selected cell or tissue type; comparing said data with data derived from the different cell or tissue type or with a control data set; and outputting at least one sequence associated with said locus or a genomic location of said active chromatin. In one embodiment, the inputted data comprises sequences of chromatin hypersensitive sites generated by enzymatic digestion of chromatin. In another embodiment, the inputted data comprises sequences of chromatin hypersensitive sites generated by using thermostable polymerase amplification of preselected regions of the genome. In one specific embodiment, the preselected regions are within 200 kb of a gene known to be associated with a disease state.

Another related embodiment of the invention provides a computer executable program for profiling a genetic locus for allelic variants affecting the formation of active chromatin, comprising inputting data comprising regions of chromatin hypersensitivity sites derived from a selected mammalian cell or tissue type; comparing said data with data derived from the same cell or tissue type isolated from another mammal of the same species with a control data set representing normal or expected sequences from said species; and outputting at least one sequence having an allelic variant affecting said active chromatin formation.

A further embodiment of the invention provides a regulatory profile platform comprising regulatory sequences associated with a plurality of genetic loci in a plurality of different cell types.

4. BRIEF DESCRIPTION OF THE FIGURES

Figure 1. Schematic illustration of an embodiment of high-throughput quantitative chromatin profiling of hypersensitive sites using quantitative PCR (HSqPCR).

Figure 2. Flowchart of an embodiment of calculating the hypersensitivity ratio from measured chromatin sensitivity data.

Figure 3. Scatter plot of HS scores for HBB K562. A baseline trend is recognizable with outliers occurring both above and below. The groups or clusters of outliers falling below the baseline are the values corresponding to candidate HS sites.

Figure 4a and b a LOWESS fitted baseline of trimmed means for HBB K562; b shows clustering of HS values represents a secondary peak to the left of the central peak.

Figure 5. Baseline with robust outlier bands HBB K562.

Figure 6. Outlier clusters identified in HBB K562.

Figure 7. Signal-to-Noise ratio scoring of the HBB K562 Locus.

Figure 8. illustrates an alignment of DNase hypersensitivity data with mouse-human conservation scores produced by AVID and visualized with rVista across the ~90kb beta-globin locus.

Figure 9. illustrates an alignment of DNase hypersensitivity data with mouse-human conservation scores produced by AVID and visualized with rVista across the T-cell receptor alpha LCR on chromosome 6.

Figure 10 depicts an illustrative approach for the assembly of DNA/Master Mix for use in qPCR reactions.

Figure 11 depicts an illustrative approach for the assembly of a qPCR reference plate.

Figure 12 a. depicts an illustrative arrangement for a re-arrayed primer plate; **b.** depicts an illustrative arrangement for a detailed qPCR reaction plate configuration.

Figure 13a-c depicts an illustrative regulatory sequence profile for the beta globin locus generated in accordance with one embodiment of the present invention.

Figure 14. a, Relative DNaseI sensitivity measurements (DNaseI-treated vs. untreated; y axis) over 25.6kb spanning the β-globin LCR (x axis: chr. 11 HG12 coordinates) in K562 cells. 783 individual measurements are shown (9 replicate determinations for each of 87 amplicons). Values are normalized to a DNaseI-insensitive reference amplicon from the inactive Rhodopsin locus. Values <1 indicate increased sensitivity to DNaseI in the treated vs. untreated sample. Evident is the average trend of clustered measurements about a baseline of fitted trimmed means (black line). 95% confidence bands are shown in orange. Measurements below the lower band are considered hypersensitive. Hypersensitive sites are rigorously identified as genomic positions with corresponding statistical outliers that cluster over replicate determinations. Means of clustered outliers are marked ('+'). **b,** DNaseI hypersensitivity expressed as computed signal-to-noise ratio (SNR) from clustered outlier data shown in **a.** HS1-5 are clearly evident. HS-7.2 identifies a previously-recognized non-erythroid-specific minor HS (Forrester et al., 1987, *Nucleic Acids Res.* **15,** 10159-77). **c,** Peaks in SNR correspond precisely with core regulatory factor binding regions of HS1-5. Shown are known binding sites for erythroid-specific regulators GATA-1 and NF-E2 (Ikuta et al., 1991, *Proc. Natl Acad. Sci. U S A* **88,** 10188-92; Strauss et al., 1992, *Proc. Natl Acad. Sci. U S A* **89,** 5809-13; Stamatoyannopoulos et al., 1995, *EMBO J.* **14,** 106-16), and the constitutive insulator-associated protein CTCF (Farrell et al., 2002, *Mol. Cell. Biol.* 22, 3820-31).

Figure 15. Identification of regulatory sequences with quantitative chromatin profiling. Peaks in SNR (vertical axes) define *cis*-active sequences relative to genomic positions/genes (horizontal axes). **a**, 66kb profile of the alpha-globin upstream regulatory region in K562 erythroid cells (K562) produced from replicate analysis of 271 amplicons (2439 independent measurements). **b-c**, Profiles of the adenosine deaminase locus (60kb; 1728 measurements over 192 amplicons) and the CD2 locus (26.2kb; 864 measurements over 96 amplicons), respectively, in T-lymphoid cells (Jurkat). **d**, 30.4 chromatin profile of the c-*myc* locus in HepG2 hepatocellular carcinoma cells (855 measurements over 95 amplicons). The profiles delineate all characterized regulatory elements in these loci including enhancers, locus control regions, and promoter elements.

Figure 16. 90.4kb quantitative chromatin profile of the human β-globin locus in K562 cells (3393 measurements over 377 amplicons). Shown (3′ to 5′) on the horizontal axis are the genomic positions of the ε-globin, $^{G}\gamma$-globin, $^{A}\gamma$-globin, δ-globin, and β–globin genes, as well as an olfactory receptor-like gene (OR5814) located 5' of the LCR. All of the major *cis*-regulatory elements of the globin locus are identified including the Locus Control Region (HS1-5); the ε-globin promoter together with an upstream element; the $^{G}\gamma$-globin promoter; the $^{A}\gamma$-globin promoter; the $^{A}\gamma$-globin 3' enhancer; the δ-globin promoter; the β-globin promoter; and the β-globin 3' enhancer. The profile identified several novel features (unlabeled peaks). Note the prominence of the δ-globin promoter, consistent with active δ-globin transcription in K562 cells (Mookerjee et al., 1992, *Blood* **79**, 820-5).

Figures 17a-d, Patterns of general DNaseI sensitivity reveal higher-order chromatin architecture in the CD2 (**a**), c-*myc* (**b**), TCR-alpha (**c**), and beta-globin (**d**) loci. DNaseI sensitivity baselines (black) and 95% confidence bounds (orange) are plotted vs. genomic position. Hypersensitive sites ('+') are situated at the epicenters of higher-order chromatin formations (marked with size in kb above). **d**, DNaseI sensitivity profiles reveal sub-domains within the ~90kb β-globin locus. Functionally active LCR and γ-gene domains exhibit increased general sensitivity in contrast to less active or inactive genes or to intergenic regions. Profiles **a-d** and of α-globin

encompassed several related and unrelated genes. However, rigid demarcation of gene domains at the chromatin level was not observed.

Figure 18 **a**, 26kb chromatin profile of the human T-cell receptor-alpha downstream regulatory region (864 measurements over 96 amplicons). Spacial organization of the TCRα 3' hypersensitive sites is thus similar to major regulatory elements of the murine locus (colored boxes), though in human an additional prominent site is present ('*'). The HS situated within the last intron of the ubiquitously-expressed *Dad1* gene was evident in several different cell types including non-hematopoietic cells (data not shown). **b**, Alignment of orthologous human and mouse sequences reveals extensive conservation across the TCR-alpha regulatory region (Koop et al., 1994, *Nat. Genet.* 7, 48-53). HS sequences identified in human T-lymphoid cells exhibit varying degrees of conservation. However, specific discrimination from other sequences in the locus exhibiting similar levels of conservation is not possible.

5. DETAILED DESCRIPTION OF THE INVENTION

The expression of a gene is coordinately regulated by numerous regulatory sequences within the gene and associated molecules. A complete understanding of gene regulation and its critical role in fundamental biological processes, including development, differentiation, and proliferation, as well as disease and other disorders, requires the identification of the regulatory sequences that coordinately control expression of a gene. The identification of such regulatory sequences and their activities in different cells or in response to different stimuli, for example, is critical to understanding and targeting gene expression, diagnosing diseases associated either directly or indirectly with gene regulation, and identifying and characterizing therapeutic protocols and drugs, for example.

The present invention provides methods for quantitative profiling of chromatin structure. The methods of the invention involves measuring a quantitative profile of chromatin sensitivity to a DNA modifying agent, e.g., DNase I. Specifically,

chromatin or cell nuclei are treated with a DNA modifying agent such that the DNA is cut or digested at appropriate sites, e.g., at sites the DNA modifying agent can access. A quantitative chromatin sensitivity profile comprising measurements of chromatin sensitivity as a function of genomic positions in a genomic region is then obtained. Preferably, the quantitative chromatin sensitivity profile comprises a plurality of replicate measurements at each of a plurality of genomic positions. The genomic positions can be represented by genomic sequences. In one embodiment, the chromatin sensitivity profile comprises a plurality of replicate measurements of each of a plurality of different genomic sequences in the genomic region. In a preferred embodiment, chromatin sensitivity at a genomic position or sequence is measured by real-time quantitative PCR as a change in copy numbers of an amplicon comprising the genomic sequence measured from chromatin that has been treated with the agent relative to copy numbers of the amplicon measured from chromatin that has not been treated with the DNA modifying agent. In one embodiment, such a change is represented by a ratio between the copy numbers measured from a treated sample and copy numbers measured from an untreated sample. For example, using real-time PCR, a highly sensitive site in the profiled genomic region, e.g., a DNase I hypersensitive site, is represented by a decrease in copy numbers of an amplicon comprising a sequence at the site measured from a treated sample relative to those of the amplicon measured from an untreated sample. Such highly sensitive sites can then be identified in a measured chromatin sensitivity profile as outliers. The invention thus also provides methods for identifying regulatory sites in a genomic locus and to methods for determining chromatin architecture in a genomic locus.

Quantitative chromatin sensitivity profiles can be obtained in vivo by treating nuclei of cells with the appropriate DNA modifying agent. Alternatively, the chromatin can be isolated from nuclei and treated with the appropriate DNA modifying agent. Any DNA modifying agent that digest DNA molecules can be used in the present invention. In one embodiment, the DNA modifying agent digest DNA molecules in a sequence nonspecific fashion. In another embodiment, the DNA modifying agent digest DNA molecules in a sequence specific fashion. Examples of DNA modifying agents include but are not limited to a non-specific endonuclease, a sequence-specific endonuclease (e.g., a restriction enzyme), a DNAse, DNase I, S1 nuclease, micrococcal nuclease, mung bean nuclease, P1 nuclease, a topoisomerase, topoisomerase II, a methylation-sensitive enzyme, *Dpn*I, *Msp*I, *Hpa*II, a chemical

DNA modifying agent, hydrogen peroxide, potassium permanganate, a DNA-modifying chemotherapeutic agent, radiation, UV radiation, histone acetylation, cytosine methylation, nuclease, topoisomerases; methylases; acetylases; chemotherapy agents that effect DNA; radiation; physical shearing; nutrient deprivation, folate deprivation, and combinations thereof.

5.1 Regulatory sequences, Regions and Profiles

The instant invention provides methods of identifying and using regulatory profiles, units and sequences, an overview of an approach is given in Figure 1. Such regulatory profiles may be directed to individual genetic loci, or they may involve characterization of multiple loci. In particular, the invention provides methods of identifying and determining the activity of the regulatory sequences associated with a gene to establish a regulatory profile for the gene. Such regulatory profiles may be used to characterize an entire gene system, including the coding and transcribed regions, as well as surrounding sequences that cooperatively regulate gene expression, as illustrated for the beta-globin locus in Figure 14

a. Regulatory or regulatory sequence profiles, as described herein, may be used to characterize a gene or genetic locus, irrespective of gene expression.

Regulatory profiles of the invention may be directed to a single genetic loci in a particular cell, but they may also encompass a plurality of loci across many different cell types and all variations between. For example, a profile of a genetic locus may be determined for one or more different cell types, and profiles of multiple genetic loci may be determined for a single cell. Furthermore, profiles of multiple genetic loci may be characterized in multiple tissues to create higher-order multiple loci profiles associated with certain cells. This would illustrate regulatory sequence profiles established for multiple loci in different cell types and depicts the unique multiple loci profiles characteristic of each cell. Furthermore, these profiles may be combined to generate a composite profile. Such composites may be useful, for example, in identifying genes or gene systems associated with certain disease indications or clinical outcomes.

Fundamentally, the invention allows the identification of regulatory sequences (RSs) associated with a gene. RSs may include multiple individual regulatory sequences and, typically, are approximately 100 to several hundred base pairs in size. Such elements and their activities may be used for a variety of purposes, including the

establishment of regulatory profiles (regulatory sequence profiles) associated with different cells.

In certain embodiments of the invention, and as described here for exemplary purposes, regulatory sequences (RSs) may be identified as hypersensitivity sites. Accordingly, RSs may include sequences, for example, present in a region of chromatin with a conformation that permits cleavage by a nuclease. RSs may be associated with particular chromatin modifications, such as histone acetylation, for example, and/or particular DNA binding proteins, such as transcriptional activators or even repressors, for example. RSs are frequently, but not always, associated with regulatory sequences capable of affecting gene expression. Even if an RS is associated with the regulation of gene expression, the affect of an RS on expression of a gene may be undetectable by current methods of examining mRNA expression, which typically require differences of at least 50% for detection. Thus, an RS that affects gene expression by less than 50% may be difficult or impossible to associate with gene expression. Furthermore, if a change or mutation of an RS is only present in one allele, its affect on gene expression is further masked by the effect of the other allele. Also, it is understood that changes in transcription may be difficult to detect by measuring steady-state mRNA levels, due, for example, to a low rate of mRNA turnover or carryover. Finally, since RSs may be formed prior to actual transcription of a gene, RSs may be detected before gene expression is actually affected.

The skilled artisan would also appreciate that RSs may also include sequences not associated directly with the regulation of gene expression, including structural components of the genomic chromatin, such as, for example, matrix attachment regions, sequences involved in the initiation or control of DNA replication, or others. Furthermore, it is understood that some RSs may form in a tissue-specific manner, for example, but may not affect transcription in their normal genomic location. For example, the erythroid specific HPFH1 enhancer forms in K562 cells but not HeLa cells and is located approximately 120 kb from the β-globin gene, where it is not thought to be active in regulating transcription. However, it was discovered due to a naturally occurring deletion, which removes the intervening sequence and juxtaposes the enhancer next to the β-globin gene, where it upregulates the β-globin gene. RSs may also include patches of randomly distributed sequence on which chromatin does

not properly form. One example is the CAG repeat structure, which does not necessarily affect gene expression.

The identification of RSs within a particular genomic region, which may include one or more genetic loci, allows the identification of regulatory systems, which function cooperatively and/or coordinately in gene expression. Furthermore, the identification of RSs associated with a particular gene or genetic locus allows further characterization of genes, *e.g.*, as coding/transcribed regions plus surrounding regulatory sequences. The identification of RSs within a certain genomic region or within a certain distance of the coding region of a gene allows the establishment of a regulatory profile for the particular genomic region or gene. Such a regulatory profile typically includes information regarding the activity of one or more RSs within the characterized region.

In certain embodiments, the regulatory profile includes information regarding the activity or sequence of a plurality of RSs or regulatory sequences within a region or associated with a particular gene. Such regulatory profiles provide important information regarding the regulation of gene expression, particularly when the regulatory profile of a gene is compared between different cells, for example. However, regulatory profiles of RSS and, in certain circumstances, even individual RSs described therein, may be used independently of any direct knowledge of their role in gene expression. For example, regulatory profiles may be established for different cells, *e.g.*, cells treated with different stimuli or disease versus normal cells, and used for diagnostic or therapeutic purposes, as described *infra*. Accordingly, it is understood that regulatory profiles and RSs of the instant invention have instrinsic value independent of any association with gene regulation *per se*, since their activity may be used as a form of genomic fingerprint to identify and characterize cells and their response to different stimuli. Furthermore, active control regions (ACRs) associated with a particular genomic location or gene may themselves be used for a variety of methods of the invention, including regulating gene expression and as probes, for example, to determine the activity of a regulatory sequence or to establish a regulatory profile of a cell. Active control regions include the sequence present with a particular region of the genome or within a certain distance of the coding region of a gene and may be isolated from the genome. In certain embodiments, an active control region includes the sequence within 50-100 kb of a gene, including all integer values in between. In related embodiments, ACRs include all sequence with

50 kb of a gene, within 60 kb of a gene, within 70 kb of a gene, within 80 kb of a gene, within 90 kb of a gene, or within 100 kb of a gene. In another embodiment, an ACR includes all sequence within 150 kb or 200 kb of a gene. In another embodiment, an active control region includes a region of between 50 and 100 kb of contiguous genomic DNA, including all integer values in between. In related embodiment, an active control includes at least 50 kb of genomic DNA, at least 60 kb of genomic DNA, at least 70 kb of genomic DNA, at least 80 kb of genomic DNA, at least 90 kb of genomic DNA, at least 100 kb of genomic DNA, at least 150 kb of genomic DNA or at least 200 kb of genomic DNA. The use of profiles including a plurality of RSs provides more opportunity to detect differences between two profiled regions or genes between different cells, since some differences might affect only one or several of the RSs located in a profiled region or gene.

A large number of active control elements, and regulatory sequences and units associated with specific loci were discovered through exploration of DNA sites in actual use by a living cell. For each identified RS, a nearby regulated gene is identified and one or more RSs or regulatory sequences associated with the gene are identified. Such RSs or regulatory sequences, alone or in combination, may be used to regulate gene expression in specific cells and at specific times, for example, and further in the identification of genes associated with a specific disease or function and the identification and development of new drugs, and novel diagnostic and therapeutic methods. Profiles may be established in any cell type and the methods of the invention describe infra may be practiced using any cell type, including for example, and not limited to, primary cell culture cells, transformed primary cells, immortalized lines, and transgenically modified lines. Cells may be derived from any lineage, including, for example hematopoietic, epithelial, liver, pancreas, brain, mesenchymal, cardiovascular, kidney, neuroectodermal, stem cell, and endothelial. Various embodiments of the invention including, amongst other things, polynucleotides comprising one or more identified regulatory units; databases and computer readable media comprising information related to the identified regulatory units, associated genes, and genomic position; and methods of using the identified regulatory units to regulate gene expression, to detect and/or treat disease, and the identification of drug candidates are described in further detail below.

Profiles of regulatory sequences may be prepared for a particular region or gene using any different type of cell. For example, cells may be isolated from

different tissues or tumors, may represent different stages of development or cell cycle; they may be treated with different stimuli, chemicals or compounds, at any concentration or for any time duration; they may be from a subject or patient with a disease or disorder; and they may be from different types of plants, microbes, or animals, including model animals such as rat, mouse, dog, pig, sheep, and primates, and mammals such as humans or mice, for example. Indeed, the skilled artisan would understand that arrays may be prepared from any type of cell. The cell or cell type may be purified or populations of different cell types may be used according to the invention.

In certain embodiments, RS profiles may be prepared for genes known to be associated with a particular biological process or disease,, as it is believed that changes in profiles are useful in detecting or analyzing the associated biological process or disease. For example, profiles may be prepared for genes known to be associated with cancer, such as, for example, oncogenes or tumor suppressor genes, including, amongst others, p53, Rb, INK4A/p16, CTNNB1, H-Ras, Fos, MDM2, INK4, ARF1, PTEN, Jun, WNT3A/14, NFkB, TERT, BRCA1, BRCA2, WAF1/p21, CDK4, TGF-beta1, RAR, E2F, VHL, MLH1, SMAD4, SMAD2, SMAD3, K-Ras, EGFR, WT1, Myc, Raf, ABL, and HER2. Other genes associated with cancer include genes involved in apoptosis, such as, *e.g.*, Bcl2, Bax, Bad, Bid, MLL, Casp3, Casp6, Casp7, Casp8, Casp9, Casp1, and BclXL.

5.2 Methods of Preparing Regulatory sequence Profiles

Profiles of regulatory sequences are generally prepared by identifying regulatory sequences within a specific genetic locus. Genetic loci are a defined region of genomic DNA. Genetic loci may include at least a portion or an entire open region frame, typically corresponding to a gene. In certain embodiments of the invention, genetic loci are selected based upon their containing a known gene. In one embodiment, known genes are associated with a particular phenotype or biological process, which may, for example, be a disease state or other disorder. The disease or disorder may be any known disease, including, but not limited to, cancers, proliferative diseases, neurological diseases, and infections. Disease-associated genes include any gene identified as playing a direct or indirect role in a disease, including, for example, the initiation, progression or biological response of a disease. Examples of disease associated genes include oncogenes and tumor suppressor genes, with

specific example being p53, Rb, INK4A/p16, CTNNB1, H-Ras, Fos, MDM2, INK4, ARF1, PTEN, Jun, WNT3A/14, NFkB, TERT, BRCA1, BRCA2, WAF1/p21, CDK4, TGF-beta1, RAR, E2F, VHL, MLH1, SMAD4, SMAD2, SMAD3, K-Ras, EGFR, WT1, Myc, Raf, ABL, and HER2.

The profiled genetic locus may include the open reading frame of the gene and sequence upstream and/or downstream of the locus. The size of a profiled genetic locus is variable and, in certain embodiments, includes greater than about 1 kb, 10 kb, 25 kb, 50 kb, or 100 kb of DNA. In particular embodiments, the genetic locus may comprise between about 1 to 100 kb, 25-75 kb, or 50-100 kb of DNA.

Methods of characterizing RSs associate with a genetic locus comprise providing a sample containing nuclear chromatin, treating the sample with an agent that induces DNA modifications in DNA with hypersensitivity sites, and identifying the DNA hypersensitivity sites induced by the agent, thereby generating an regulatory sequence profile associated with the genetic locus. Depending on the accessibility of the site to the agent, more or less cleavage will occur, allowing quantitative analysis of the activity of RSs.

The invention contemplates the use of a variety of methods to identify and analyze RSs within a genetic locus, for which exemplary procedures are provide in the accompanying Examples and described in U.S. patent applications No. 60/108,206, No. 09/432,576, No. 60/302,369, No. 60/290,036, No. 60/294,890, No. 60/294,890, No. 60/378,664, No. 60/387,910, No. 60/387,887, No. 10/187,887, and No. 60/404,121, and PCT applications PCT/US02/15032 and PCT/US02/16967 are specifically and entirely incorporated by reference.

In one embodiment, quantitative PCR methods are used to identify hypersensitivity sites and corresponding RSs within a genetic locus. Where the sequence of at least certain regions of the genetic locus is known, PCR primers may be designed to amplify amplicons comprising a portion or the entirety of the genetic locus. In certain embodiments, PCR primers may be used that produce overlapping or adjRSnt amplicons, which include substantially all or the entirety of the sequence corresponding to the profiled genetic locus.

In certain embodiments, the goal of determining a profile is to identify regulatory sequences, for example, DNAse I hypersensitivity sites, within and surrounding a genetic region of interest. Accordingly, locus profiles may be designed to cover the promoter, the first few introns and immediately 3' of the last exon of a gene under the

presupposition that these segments are the most likely to contain regulatory elements and exert control over gene expression. The size of a locus profiling experiment depends, in part, on the size of the gene or gene cluster being analyzed. A locus profile of approximately 250 amplicons may typically span 20-50 kb. By performing quantitative PCR within each amplicon, a loss in copy number, resulting from digestion with DNase I, for example, can be reproducibly detected.

The activity of an RS is typically calculated based upon copy number determined by quantitative PCR, with higher copy number indicating lower activity of the RS. The quantitative PCR data may be converted into scores for each amplicon and the values plotted versus the genomic position to yield a DNAse hypersensitivity graph or RS profile.

In one illustrative example of how locus profiling data may be represented, the hypersensitivity site score is determined in a relative fashion by comparing the copy number of each target amplicon to a reference. Reference amplicons may be selected from genes that are not expressed at an appreciable level in the cell type being examined by the locus profiling experiment, When a gene is not expressed, the chromatin is believed to be in a closed conformation, In such a case, DNase I does not have easy access to the DNA and cannot digest it when it is wound around nucleosomes. The reference amplicon allows us to estimate the copy number of a DNA sample at a site that it not susceptible to DNase I digestion.

5.3 Methods Using Regulatory sequence Profiles

Regulatory sequences and profiles thereof may be used in a variety of different applications and methods according to the invention, as will be understood by the skilled artisan. Embodiments related to a variety of methods using regulatory sequences and regulatory units are described *infra* for exemplary purposes, and the skilled artisan would understand that such methods may also be adapted and performed using regulatory sequence profiles instead of individual regulatory sequences or active chromosomal elements or units comprising a plurality of active chromosomal elements. The adaption of the described methods to using regulatory sequence profiles would require merely routine procedures. Furthermore, it is understood that any of the method described herein may be performed using high throughput techniques, including, for example, the use of microarrays and robotics.

Examples of further methods directed to the use of regulatory sequence profiles are provided below.

As described above, regulatory sequence profiles may be established for any cell type and any conditions. Therefore, regulatory sequence profiles may be generally used to identify or characterize a cell. In one embodiment, an regulatory sequence profile may be established for a cell of unknown type, *e.g.*, unknown tissue origin, or unknown if diseased or not. This profile may then be compared to profiles of known cell types to identify the type of the unknown cell. In certain situations, the regulatory sequence profile associated with one genetic locus may be sufficient to characterize a cell, for example, as having a particular disease or tissue origin. However, in other circumstances, a plurality of different regulatory sequence profiles associated with different genetic loci may be used to identify a cell. Thus, in different embodiments of the invention, methods of identifying or characterizing a cell are provided which compare regulatory sequence profiles of one or more different loci. In certain embodiments, the comparisons are performed using arrays or microarrays comprising polynucleotide sequences corresponding to regulatory sequences associated with one or more genetic loci. In another embodiment, regulatory sequence profiles used for comparison may be established using PCR primers specific to one or more genetic loci.

The methods of identifying cells based upon their regulatory sequence profiles may be used for a variety of purposes. For example, the cell type of a cell of unknown tissue origin may be determined. This is particularly advantageous, for example, to determine the tissue origin or cell type of a metastatic tumor or circulating tumor cells.

In related embodiments, these methods may be used in disease detection and diagnosis. An regulatory sequence profile associated with a specific disease or disorder may be determined, for example, by comparing the regulatory sequence profiles associated with one or more genetic loci between a normal and a diseased cell and identifying a difference. Once a locus having a different regulatory sequence profile in a diseased cell is identified, this information may be used for diagnostic purposes. For example, an active regulatory element profile associated with the identified locus may be prepared or identified for a cell suspected of having the disease and compared to the profile of a diseased cell and/or a normal cell. If the profile of the suspect cell is the same or substantially similar to the profile of the

disease cell, then the suspect cell is considered to have the disease. In certain embodiments, a profile may be considered to be the same if the measurable activity of all of the identified regulatory sequences within the analyzed locus are within an approximately 10% or 25% range of the activity in the diseased cell. A profile may be considered to be substantially similar if the activity of a majority of the regulatory sequences within the analyzed locus are within an approximately 10% or 25% range of the activity in the diseased cell. It is understood that the disease profile may involve a change in activity of a single regulatory sequence as compared to a normal cell, or it may involve changes in the activity of more than one or a plurality of regulatory sequences. Therefore, it is understood that if a single regulatory sequence within a profile is indicative of disease or cell type, then it is the activity of this element that is relevant to the determination of cell type or disease state. Accordingly, a profile may be substantially similar if the activity of an indicative regulatory sequence is closer to that observed in one cell type as compared to another.

Profiles of the invention may also be used to characterize cells for a variety of other purposes. For example, profiles may be established that are associated with different clinical outcomes for a particular therapy, for different responses to drug treatment, and for characterizing patients in a clinical trial. Accordingly, the invention provides methods of predicting clinical outcome of a therapy that involve identifying a particular profile associated with a clinical outcome and comparing the profile of a patient to this profile to correlate and predict the clinical outcome of a particular therapy for a specific patient. Similarly, the invention provides methods of determining whether to use a drug to treat a patient, which involve identifying a profile associated with either an adverse or positive outcome of drug treatment and comparing this profile to the corresponding profile in a patient to determine whether treatment with the drug is appropriate. Such methods are particularly valuable in identifying the potentially small number of patients who may have adverse effects to a drug.

Profiles may also be used to identify a gene associated with a disease or disorder. For example, profiles for one or more genetic loci may be compared between a normal and disease cell, and profiles specifically associated with a disease or disorder identified. Once a profile and corresponding loci associated with a disease or disorder is identified, the gene associated with the disease or disorder may be identified based upon its physical proximity to the loci. For example, if the loci

encompasses an open reading frame, the gene associated with the disease may be associated with this open reading frame. In addition, the gene associated with the disease may be located within a certain distance of an regulatory sequence showing different activity in the disease cell, for example, within 1, 2, 5, 10, 20, 50, 100 or 200 kb of the element. Open reading frames and genes located within a specific chromosomal region may be readily determined by methods available in the art, *e.g.* using publicly available genome databases. In certain embodiments, profiles may be prepared for one or more known gene loci, including, for example, genes suspected or known to be associated with a disease.

Profiles may also be used to identify allelic forms of a gene. For example, an regulatory sequence profile associated with a gene in one cell (reference cell) may be compared to an regulatory sequence profile for the same gene in one or more other cells. In certain embodiments, the same cell type is used to reduce the possibility that any differences in profiles are due to cell type, *etc.* rather than gene variation. Cells displaying a different profile than the reference cell may be identified and the sequence of regulatory sequences displaying different activities determined to identify allelic variants. The invention further provides allelic forms of a gene or allelic variants identified by such methods.

Profiles may further be used to identify a regulatory element or sequence active in a particular cell. For example, profiles prepared using different cell types for a particular genetic loci may be compared to identify an RS that has increased activity in a particular cell type, thereby indicating that the identified RS corresponds to regulatory sequences active in the cell. Such sequences may be used, for example, to direct expression of a linked gene to the particular cell.

In certain embodiments, RS profiles may be used to screen candidate drugs to identify a drug useful for treating a disease or disorder. A profile associated with a disease or disorder may be identified according to methods of the invention, for example, by identifying a profile characteristic of a diseased cell and not a normal control cell. A disease cell may then be treated with a candidate drug, and the drugs affect on the cell examined by determining whether the profile associated with the disease changed and became more similiar to the profile associated with the normal control cell following treatment with the candidate drug. Drug candidates that cause a disease-associated profile to become more like the normal cell profile are thus identified as candidates that may be useful in treating the associated disease. The

invention contemplates screening any and all known and available drug candidates and types of molecules, including, for example, polypeptides, antibodies, polynucleotides, hormones, cytokines, organic and inorganic molecules, and small molecules. The invention further includes drugs identified by a method of the invention, and methods of manufacturing a drug comprising identifying a drug that alters an RS profile associated with a disease or disorder and manufacturing the identified drug.

5.4 Methods of determining chromatin sensitivity profile

Real-time PCR based quantitation of chromatin sensitivity can be carried out by the methods described in PCT publication WO 02/097135 and McArthur et al., 2001, J Mol Biol 313, 27-34, each of which is incorporated herein by reference in its entirety.

In a preferred embodiment, DNaseI is employed as the DNA modifying agent. DNaseI digestions can be carried out according to standard protocol (see, e.g., Reitman et al., 1993, *Mol. Cell. Biol.* 13, 3990-8). The amount of DNaseI for each reaction and the DNaseI incubation time can be determined by one skilled person in the art.

Following DNaseI treatments, DNA can be purified according to standard protocol known in the art, e.g., using the Puregene system (Gentra Systems, Minneapolis, MN) according to the manufacturer's protocol. The purified DNA is then resuspended in 10mM Tris-Cl, pH 8.0. In one embodiment, DNA samples are quantitated in triplicate using a Spectramax 384 Plus UV spectrophotometer (Molecular Devices Corporation, Sunnyvale, CA).

In a preferred embodiment, purified genomic DNA prepared from DNaseI-treated and untreated chromatin samples are used as qPCR templates to determined the DNaseI sensitivity profile. A plurality of primers are used to amplify a plurality of sequences, *i.e.*, amplicons, across the genomic region of interest. The plurality of sequences can be successively overlapping and therefore covering the entire genomic region without gaps. Alternatively, if desired, the plurality of sequences can consist of non overlapping sequences. The size of each sequence in the plurality of sequences can be the same or different. The sizes of the amplicons can be determined by one skilled person in the art based on the desired density of coverage of the target genomic region, which in turn depends on factors such as the distances between

different hypersensitive sites and so on. Allowing different amplicon sizes facilitates selection and design of primers. Preferably, the sizes of amplicons can range from about 75 bases to about 300 bases. In a preferred embodiment, primers are selected such that amplicons having a mean size of 250 bases are amplified. To increase the density of coverage over a genomic region, amplicons having 50, 100, 150 or 200 bases can be used. To decrease the coverage over a genomic region, amplicons having 280 or 300 may be used. It will be apparent to one skilled person in the art that for a given genomic locus, different density of amplicons can be used for different regions. In a preferred embodiment, primers are designed to amplify successively overlapping amplicons across a target genomic region.

Primers can be designed using standard method known in the art, e.g., Primer3 (Rozen, S., Skaletsky, H.J. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz, S. and Misener, S. Bioinformatics Methods and Protocols: Methods in Molecular Biology, Humana Press, NJ, 2000) restricting several parameters, including target amplimer size (250 bp +/- 50 bases); primer T_m (optimal, 60°C +/- 2°C); %GC (50% optimal, range 40-80%), and length (optimal 24, range 19-27); and the poly X (maximum 4). Preferably, primer design parameters are optimized and calibrated empirically to navigate as efficiently as possible anomalous features of the genomic terrain such as repetitive elements and purine or pyrimidine isochores. In one embodiment, primers are scanned for repetitive sequences by BLAST alignment with the Alu and NR databases. In another embodiment, a highly effective mispriming exclusion algorithm based in part on a comprehensive positional 16-mer index of the human genome is used (see, e.g., U.S. Patent Application No. 10/375,404, filed February 27, 2003, by Stamatoyannopoulos et al., which is incorporated by reference in its entirety).

In a preferred embodiment, the sensitivity of each amplicon to DNaseI digestion is measured by quantifying relative copy ratios between DNase-treated and untreated samples. In order to arrive with a common DNase sensitivity scale that could be applied to other loci, relative copy ratios can be normalized to a standardized reference amplicon from an appropriate gene locus, *e.g.*, a locus which is transcriptionally inactive and DNaseI-resistant in the cell type.

In a preferred embodiment, a plurality of replicate measurements are performed for each amplicon. Using the replicate measurements, a relative DNaseI sensitivity profile is constructed. A statistical bootstrap validation approach was

carried out over replicate data sets to evaluate the reproducibility of individual experimental data and also to identify the degree of replicate coverage required to achieve high accuracy. Bootstrap sub-sampling was applied to highly replicate (9X) data sets; varying numbers of HS measurements per genomic location were evaluated with a goal of determining whether sites identified in the highly replicate dataset could be detected using a subset of the data. This analysis provided convincing results that even at lower replicate coverage levels accuracy remains high. Typically, approximately 93% of the HSs identified in a 9X replicate set were identified in low (3X) replicate sets while retaining excellent average specificity. At the 6X replicate level, all of the sites identified at the 9X level are generally detected. Sites with higher signal-to-noise ratios > 7.0 and consistently low measurement error (< median baseline variance) are consistently detectible (variance 0.001) whereas the variation in false positives for 6X replicate bootstrap is still only 1%. This provides strong evidence that smaller replicate (3-6) datasets having modest measurement error with respect to the baseline provide sufficient data for accurate HS detection. Thus, in preferred embodiments, 3 to 6 replicate measurements are carried out for each amplicon. More replicate measurements can also be carried out, e.g., 9 or more, if desired.

Quantitative PCR (qPCR) can be carried out according to standard protocol. In one embodiment, 15μL real-time quantitative PCR reactions are prepared using 0.9μM forward and reverse primers, 30 ng template DNA (untreated or DNaseI-treated) and master mix composed of 1X FastStart buffer (Roche), 200μM of each dATP, dCTP, dGTP, dTTP, 3mM $MgCl_2$ and FastStart Taq DNA polymerase (0.033 U/μL). The reaction mixture is supplemented with 0.33X SYBR green I stain and 300 nM 6-ROX (Molecular Probes, Eugene, OR) to detect the accumulation of PCR product during amplification and normalize fluorescence intensity, respectively. Preferably, samples are run in triplicate on individual 384-well plates, and thermalcycled with an ABI 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA).

qPCR reactions are performed using a suitable detection instrument (*e.g.*, the ABI 7900HT Sequence Detection System) as follows. Nuclei are isolated and treated with DNaseI under a standard protocol, following which the genomic DNA is purified for use as a qPCR template. Primers are designed to amplify a study region and the

progress of the amplification is followed by measuring the accumulation of signal from the double stranded DNA-specific fluorescent dye SYBR green. The instrument measures this increase in fluorescence as a function of cycle number and from the resultant amplification profile calculates the number of cycles of PCR needed to amplify the product above a specified threshold. Every sample tested is run synchronously with a standard curve so that copy number can be calculated. The number of copies of the test amplicon are normalized with the number of copies of a (similarly sized) reference amplicon designed to a DNaseI-insensitive region of the genome. A quantitative digestion profile is generated by calculating the copy loss of a test amplicon across a series of DNaseI digestion conditions and expressing this loss relative to the number copies of the DnaseI-insensitive reference amplicons (set at 100%).

5.5 Method for correction of amplification efficiency

The performance of real-time quantitative PCR in a high-throughput format is critically dependent on reliable correction for variant amplification efficiencies between amplicons of different length and base composition. In standard practice, amplification efficiencies are typically derived from the log-normalized slope of a standard curve. This approach would require that a standard curve be produced for each amplicon tested, which would in turn substantially increase the cost of profiling while reducing its throughput. Accordingly, in one embodiment, the invention provides a method for automated correction of amplification efficiency based on the use of a single standard curve.

The method can includes four steps: 1) determination of cycle threshold (Ct), 2) amplification efficiency correction, 3) melting curve analysis, and 4) calculation of hypersensitivity ratios (HRs).

After the reactions are complete, the normalized fluorescence data are exported using the ABI SDS software and then analyzed using software that determines the starting copy number for every target amplicon, *i.e.*, amplicon from the treated sample, relative to the corresponding reference amplicon, *i.e.*, amplicon from the untreated sample. In one embodiment, the exported fluorescence data are used to generate an amplification plot and an Nth-order polynomial fit of the amplification curve calculated for each reaction. At this point, an amplification threshold is set for the entire plate and the Ct values determined for each amplification curve. It is

critical to the assay to establish the amplification efficiency for each amplicon tested on a plate so efficiency differences between the test and reference amplicons can be corrected. Uncorrected efficiencies will result in erroneous copy number estimations.

In one embodiment, amplification efficiency of each reference amplicon is determined empirically for every reaction plate using a standard dilution series of DNA and the equation $E = 10^{-1/slope}$ (Liu et al., 2002, *Anal. Biochem.* 302, 52-9). The efficiencies of each of the test amplicons is determined by examining the slope of the linear portion of the amplification curve, similar to the method proposed by Ramakers et al., 2003, *Neurosci Lett.* 339, 62-6. Efficiency corrections are then performed on all test amplicons with respect to each of three reference amplicons (independently).

In one embodiment, melting curve analysis is performed. Because SYBR green I detects all double-stranded products melting curve analysis must be performed for each test amplicon. Different DNA fragments will typically possess unique melting behaviours, thereby allowing rapid identification of reactions that contain more than one product. A reaction that yields multiple products or primer dimers will yield erroneous HR values due to competition for reagents during amplification, especially if different products possess different amplification efficiencies. It is therefore required in this assay that only one product is amplified and all melting curves are scored for multiple products. Melting curve analysis is conducted for each amplicon to discard those yielding multiple products.

In one embodiment, following efficiency correction, the comparative Ct method (Livak *et al.*, 2003, *Methods* 25, 402-8) is employed to calculate relative copy number differences. Use of the comparative Ct algorithm for relative quantification is critically dependent on the assumption that reference and test amplicons have equal amplification efficiencies. Efficiency-corrected Ct values are used to compute a relative copy number ratio by applying the algorithm: $2^{-\Delta\Delta Ct}$ or $2^{-[\text{treated (target – reference)} - \text{calibrator (target-reference)}]}$ (Livak *et al.*, 2003, *Methods* 25, 402-8). The comparative Ct calculation yields a hypersensitivity ratio (HR) (=relative copy ratio). A ratio of 1 implies that there is no difference in relative copy number between the test and reference amplicons in the treated DNA sample examined. A score of < 1 is indicative of copy loss due to enzymatic cleavage with DNase I. For example, a score of 0.5 would indicate 50% copy loss relative to the reference, or DNase I insensitive amplicon.

5.6 Methods of identifying regions of chromatin sensitivity

The invention also provides a method for scoring measurements resulting from ratio-based data such as arise in DNA microarray and locus profiling applications and an overview is given in Figure 2. The method has merit in that it makes few distribution assumptions about the data and is robust under a relatively broad set of profiling scenarios while allowing significant replicates to be visualized. The main source of subjectivity in the model is in the fitting of the baseline as it may depend on the large-scale chromatin structure of the locus. The model is applicable to a wide set of initial conditions and data sources.

The advent of real-time quantitative PCR (qPCR) and recent developments in improving its robustness enables large scale genomic profiling and genotyping. A key determinant in the process is quantifying the amplification efficiency, and common methods are based on measuring an amplification ratio using the method of standard curve (see, e.g., Ramakers, C. Ruijter, J.M., Deprez, R.H.L. Moorman,A., *Assumption-free analysis of quantitative real-time polymerase chain reaction (PRCR) data,* Neuroscience Letters 339 (2003) 62-66, in particular p.62, and Meijerink, J. Mandigers, C. van de Locht, L., Tonnissen, E., Goodsaid, F., and Raemaerkers, J., *A Novel Method to Compensate for Different Amplification Efficiencies between Patient DNA Samples in Quantitative Real-Time PCR,* Journal of Molecular Diagnostics, Vol 3, No.2, May 2001, 55-61, in particular pp.56-57). In this method, a serial dilution of a known quantity of DNA is typically used to calibrate the sample of unknown quantity and ratios of PCR efficiency are examined. The data generated by this method are ordinarily ratios of repeated measurements in which both numerator and denominator are subject to assumed measurement error. Much work has been done in attempts to remove some of the distributional assumptions (Brody *et al.*, 2002, *Proc. Natl. Acad. Sci. USA* 99:12975-12978) and improve the methods of relative quantification. Problems such as different accurate measurement of copy number, and interpretation of results are some of the key issues.

In the present invention, qPCR is used to measure DNA digestion by a DNA modifying agent, e.g., the enzyme DNase I, to accomplish large scale genomic profiling. The method relies on measurements of the difference in amplification yield between a control genomic DNA sample derived from untreated nuclei and one or more experimental samples from nuclei treated with varying concentrations of

DNase I before preparation of genomic DNA. A plurality of amplicons covering a genomic region of interest are measured by, e.g., qPCR, in the treated and untreated samples. Preferably, the amplicons are closely spaced along the genomic locus and covering as much as possible of the region of interest. For example, if a DNaseI cut occurs within a sequence covered by an amplicon, the amplicon will not be successfully amplified and there will appear to be fewer copies of the amplicon in the DNaseI treated sample than in the untreated control sample. The measurements of DNAse hypersensitivity in this method take the form of a ratio of copy loss between the reference and experimental samples and indicate cutting in the region of a particular amplicon. Regions of higher DNAse hypersensitivity are indicated by lower values of the copy loss ratio, and repeated analysis produces a distribution of hypersensitivity scores (HS) for the amplicon.

The qPCR assays of the series of contiguous and neighboring amplicons produces a profile of the hypersensitivity and chromatin structure of a given genomic locus comprising measurements of chromatin sensitivity, e.g., DNase hypersensitivity, as a function of genomic positions (see, *e.g.*, Figure 3). Preferably, the profile comprises a plurality of replicate measurements at each of the genomic positions. While the HS scores can be accurately measured for most amplicons (exceptions lie in exceptionally high G+C regions which must be treated separately), the copy loss of DNAse cutting is relative to the local chromatin structure in the location of the amplicon and hence HS scores are relative. There is a baseline response to copy loss in a region, and it is the deviation of repeated measurements from this baseline that is of interest in quantifying. In one embodiment of the invention, a score is given to characterize the deviation. Preferably, the score is a continuous, statistically valid; score that measures the relative intensity or significance of HS values with respect to the average chromatin profile of the locus. Chromatin sensitive sites, e.g., DNase HS sites, are then identified based on the score.

The invention provides a method for identifying chromatin sensitive sites, e.g., DNase HS regions. Figure 3 shows the scatter plot from a series of replicate measurements of HS values for a series of amplicons in the vicinity of the β-globin gene in the cancerous cell line K562 under 8 units treatment of DNase (HBB K562). Evident in the locus profile is an average trend of clustered HS measurements at an average value of about 0.75 with outliers occurring both below and above this region.

An important observation from this scatter plot is that many of the outliers below the baseline are reproducible in the sense that they can be seen to cluster reasonably tightly about particular lower values, while other outliers above and below the baseline are the product of more random noisy measurements. These values may occur due to failed primer reactions and a variety of other laboratory conditions. If these clusters can be reliably identified then a set of putative HS sites can be found.

Preferably, the method involves the following steps:

1. Recognize the trend or baseline behaviour of the locus.
2. Determine the measurement error for data clustered around the baseline, and hence empirical confidence bounds on outliers and extreme values.
3. Identify outliers that have clustering behavior or low variance with respect to the mean measurement error, eliminating isolated values and others from consideration. Examine contiguous regions of outlier clusters for possible extended HS structure
4. Assign a signal-to-noise ratio (SNR) and/or P-value to quantify the significance of this observation from the baseline. Adjust scores for contiguous structure.

Determination of the Baseline

An important observation that recurs throughout the analysis is the non-Gaussian behavior of measurement of the distribution of HS scores, and special means are taken to address this issue. The ratio x/y of two measurements each assumed to have Gaussian error term in not be distributed as a normal random variable. For small variance of the measurements (on the order of less than the mean value) in both the numerator and denominator, the ratio of observations follows a Gaussian distribution. However as the standard error increases, the ratio of measurements from Gaussian random variates approaches the Cauchy or Lorentz distribution. This has been demonstrated to be the case in particular in the analysis of DNA microarray data (Brody *et al.*, 2002, *Proc. Natl. Acad. Sci. USA* 99:12975-12978) where more robust methods for treating outliers are often necessary.

The HS values that result from repeatedly profiling a fixed region or locus exhibit an average DNase sensitivity in that region, and the initial goal is to detect that trend. In one embodiment, an initial single pass of the data is made to remove egregious outliers, e.g., data points due to multiple PCR amplification products.

Typically these will be ratios with zero or near zero values (< 0.05) or very large (>2.0) values. In embodiments in which the clustered behavior below the baseline is to be evaluated, the truncation point for the larger values is not critical.

In a preferred embodiment, a linear pass is then made through the dataset applying a suitable percent trim to the plurality of replicates measured for each amplicon. In preferred embodiments, a linear pass is then made through the dataset applying a chosen % trim, e.g., 20% trim, to the plurality of replicates measured for each amplicon. For a modest number of amplicon replicates, e.g., 3-10 replicates, this removes the most significant remaining deviates from the bulk of the data centered on the baseline. The remaining data is then smoothed. An optimal smoothing algorithm in this context is one that allows for significant local variation in the data, non-specified functional form, few parameters. In a preferred embodiment, the smoother Locally Weighted Least Squares (LOWESS) is employed to smooth the data (see, *e.g.*, Cleveland, 1979, *J. Amer. Statistical Association* 74: 829-836). LOWESS is based on robust locally-weighted regression fitting of low degree polynomials to each point using a local environment of the data. The amount of local data to include for the least squares fit at each point is conventionally determined by the tri-cube weight function as proposed by Cleveland.

$$w(x) = \begin{cases} (1-|x|^3)^3 & |x|<1 \\ 0 & |x|>1 \end{cases} \qquad (1)$$

Specifically, in embodiment the smoothing is performed by considering all the data replicates at a given genomic position and using equation (1) defined on the unit interval [0,1]. The data from five (5) neighboring amplicons, i.e., genomic positions, are used on each side of a given amplicon x to be locally smoothed. The above function (1) is mapped linearly so that local value x has w(x) = 0, while w(x-5) = w(x+5) = 0, so that the weights go to zero at this point. The value of w(x) explicitly determines the number of data points used at the amplicon value x in the local fit. A standard reference for this algorithm can be found in (Chambers et al., *Graphical Methods for Data Analysis*, Wadsworth 1983) and implementation can be found in the statistical programming languages S-Plus/R. When the degree of the local polynomials (linear) has been chosen, a single parameter $f \in (0,1)$ controls the size of the local smoothing window. In most applications of scatter plot smoothing this value ranges from (0.15,0.5) with smaller values capturing more variation in the data. In a

preferred embodiment, a value of 0.2 is used. The overall algorithm is robust to minor variations in the fitting at this stage, and there is more loss of information due to under rather than over fitting. An example of a smoothed baseline for HBB K562 is given in Figure 4a.

Centering the data about the LOWESS determined baseline yields a better understanding of the distribution of HS scores around the baseline. The clustering of HS values represents a secondary peak to the left of the central peak as shown in Figure 4b.

Determination of the Error Bounds for the Baseline

The next step is quantifying the noise about the smooth baseline so that outliers can be effectively recognized. In one embodiment, the replicate measurements for each genomic position are first mean centered about the moving baseline to generate a mean-centered chromatin sensitivity profile. The centered data are then analyzed as described in the following. The outliers of this distribution are determined using a median absolute deviation approach that is robust to finite sample breakdown. As the HS scores are derived from the ratios of measurements, care must be used in determining outliers, since for a standard normal random variable 99% of the mass is between –2.58 and 2.58, while for a Cauchy C(0,1) random variable the same mass is contained within –63.66 to 63.66.

For a Cauchy distribution $C(\mu,\sigma)$ with probability density function given by the equation

$$f_X(x) = \frac{1}{\pi}\frac{\sigma}{\sigma^2 + (x-\mu)^2} \qquad (2)$$

the moments of any order do not exist. However, robust point estimators of location are available and we have $\hat{\mu} = MED(n)$, the sample median of the observations, and $\hat{\sigma} = MAD(n)$, the median absolute deviation, and n is the number of data points. The sample median M of the data D is defined in the usual manner as $M = X_{(m)}$ where $m = (n+1)/2$ if n is odd, and $M = (X_{(m)} + X_{(m+1)})/2$ if m is even. The Median Absolute Deviation (MAD) is defined as the median of the data set $|X_i - M|$ where $X = \{X_i\}$ is the data and M is the median.

A variety of rules are available based on various distributional assumptions. In one embodiment, the MAD is used as the measure of scale for a Cauchy distribution. Therefore, data that lie a significant distance from the sample median in units of MAD are discarded. In one embodiment, the method of Rouseeuw and van Zomeren (Rousseeuw *et al.*, 1991, *J. Amer. Statistical Association* 85: 633-639) is used to declare a data point X an outlier if

$$\frac{|X-M|}{MAD/0.6745} > 2.24 \qquad (3)$$

where M is the sample median and MAD is the average median deviation. The factor 0.6745 is a correction factor for comparing non-normally distributed data, and the factor 2.24 arises in details concerning the outlier masking. Specifically, robust estimates of location and scale are used in the calculation of the Mahalanobis distance resulting in a robust measure of distance.

The procedure in this step of the algorithm is to compute outliers at each genomic location rejected using this rule, and then to define lower and upper confidence limits on the remaining data as the minimum of the upper outlier boundary, and the maximum of the minimum outlier boundary. Trimming the data in this way removes both the lower and upper extremes of the distribution in a manner that it addresses the problems of masking due to low sample breakdown.

In other embodiments, a bootstrap method is applied to determine outliers. In one embodiment, a series of bootstrap replications are performed and method is as follows:

a) At each genomic position randomly selecting one data point, i.e., selecting one replicate measurement among the plurality of replicate measurements of the genomic position, defining this dataset to be a bootstrap sample. Preferably, the data point selected will not be an outlier and will be representative of the central distribution. The bootstrap sample represents measuring HS values from a single pass of the qPCR system on the locus.

b) Performing the outlier rejection test of Rouseeuw and van Zomeren (Rousseeuw *et al.*, 1991, *J. Amer. Statistical Association* 85: 633-639) on this bootstrap sample, and determining the maximum lower outlier and minimum upper outlier values.

c) Repeating steps a) and b) for a plurality of n times and computing the upper and lower outlier cutoff values and BCa confidence intervals. Preferably, n is

at least 100, 500, 1,000, or 10,000. An ordinary skilled person in the art will be able to determine the desired value of n based on, e.g., the number of genomic positions and the number of replicate measurements in the chromatin sensitivity profile. The 100%(1-α) Bca confidence interval is a bias corrected accelerated percentile interval and is standard in the theory of bootstrap statistics (see, e.g., Efron, B. and Tibshirani, R.J., *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman and Hall/CRC 1993).

The maximum of the lower outliers and minimum of the upper outliers are obtained in this way and this provides independent constant lower and upper boundaries for the outliers of the baseline. For dense data sets involving > 75% of the data clustered around the baseline, a very small number of bootstrap replicates are sufficient. Figure 5 illustrates the results of determining the lower and upper confidence bands.

The bootstrap method is particularly useful for sparse data sets. For example, the bootstrap technique provides a highly accurate characterization of the outlier confidence band for fewer than 4-5 replicates per genomic position. Therefore, in one embodiment, the bootstrap method is preferably used when there are about 4-5 or less replicate measurements per genomic position.

Classifying Outliers for Scoring

Clustered events that are outside of the noise threshold from the baseline are then identified. In one embodiment, another linear pass of the data is performed for identifying groups at a common genomic position whose 20% trimmed mean lies strictly below the interpolated value at the lower shifted baseline. Trimming data using other percentage value can also be used. These represent events for which there is a statistically significant cluster of values that lie sufficiently below the lower outlier baseline so as to represent chromatin sensitivity at that particular locus. A small correction factor eliminates from consideration groups with very high variance or those consisting of a single point (zero variance): isolated points are immediately eliminated from consideration, those with variance strictly greater than the average variance of the baseline are also eliminated. The remaining events are termed *scorable e*vents. In one embodiment, clusters of HS values failing to meet the above criteria but bordering on scorable events are considered for missing data or failed

primer reactions and may be smoothed over rather than simply failing to be scored. Figure 6 illustrates this step.

Scoring Hypersensitivity

The deviation from the average chromatin profile, *i.e.*, the baseline, of a locus is then scored. The standard statistical approach to scoring P-values against approximations to normal distributions has been successfully used in a variety of genomic applications. In one embodiment, a p-value is calculated based on Cauchy distributions. The P-value for the cluster assuming a Cauchy distribution is easily derived from the observed information using standard techniques (see, e.g., Casella, G. and Berger, R.L., *Statistical Inference*, Duxbury Advanced Series, Wadsworth Group, 2002) and leads to a test statistic $Z = \sqrt{n/2}(HS_i - B_i)$ where the one sided null hypothesis is $H_0 : HS_i = B_i$ against $H_0 : HS_i <= B_i$. The Z statistic is well known to asymptotically approach a normal distribution with 0 mean and unit variance. These methods can be carried out with the S-plus/R statistical packages.

In another embodiment, a signal-to-noise (S/N) ratio is calculated for the locus. The S/N ratio can be calculated according to the equation

$$S/N_i = \frac{|HS_i - B_i|}{MAD_B(\sigma_c / \sigma_{HS})^2} \qquad (4)$$

where S/N_i, the signal-to-noise ratio at site i is measured as the absolute deviation of the trimmed mean (e.g., 20% trimmed mean) of the corresponding HS cluster, HS_i, from the interpolated baseline, B_i, divided by the median average deviation of the centered baseline, MAD_B. The remaining term $(\sigma_c / \sigma_{HS})^2$ is a small correction factor that penalizes larger variances in HS clusters and rewards highly compact clusters that are strongly indicative of HS sites. The factor σ_{HS} is computed as the average variance of an HS cluster of data, that is, the data assigned to an HS scorable site as determined by the algorithm. The factor σ_c is the variance of the data in the particular HS cluster being scored. It is simply the ratio of the variance of the data comprising the HS cluster to the average variance of data assigned to HS clusters computed over all scored data.

As there is noise associated with both the baseline and the HS cluster, in still another embodiment, a modified Welch two-sample t-test (see, e.g., Wilcox, Rand R.

Applying Contemporary Statistical Techniques, Academic Press, 2003) is used for comparing heteroscedastic groups. The Welch two sample t-test tests the hypothesis of equality of means subject to possibly distinct but known variances of two sample populations. It can be calculated in any of the common statistical packages available.

An example of the result of scoring the HBB locus with SNR is shown in Figure 7. It can be verified to accurately score all of the known hypersensitive sites in the HBB locus. Hypersensitive sites can be identified based on the scores. In one embodiment, the hypersensitive sites are identified if the score is above a given threshold.

In one embodiment, the invention also provides a method of contextualizing HS elements on a quantitative basis relative to one another, to their immediate flanking regions, and to their chromosomal domains generally. The chromatin profiles reveal the presence of numerous prominent perturbations representing zones of significantly increased sensitivity extending over the covered genomic region.

Although in this section the method is described in the context of identifying chromatin hypersensitivity, it will be apparent to one skilled person in the art that the method is equally applicable for identifying genomic sites where loss of sensitivity to a DNA modifying agent, e.g., DNase, occurs. These sites correspond to outliers above the baseline.

5.7 Computational analysis of quantitative chromatin profiles

The invention also provides methods for computational analysis of the quantitative chromatin profiles. The methods can be used to determine the correspondence between HSs and evolutionarily conserved non-coding sequences.

One surprise from the recent analysis of the mouse and human genomes is the relatively large portion of the mouse genome that is evolutionarily conserved but does not code for proteins (Mouse Genome Sequencing Consortium, 2002). Presumably, much of this non-coding DNA regulates the rate at which individual genes are transcribed. Therefore, it is desired in analyzing the HS data is to determine the extent to which HS's correlate with conserved non-coding sequences. In one embodiment, this comparison is carried out using rVista (Loots *et al.*, 2002 *Genome Res.* 12; 832-9) and the Genome Browser Database at UC Santa Cruz (Kent *et al.*, 2002 *Genome Res.* 12; 996-1006). Figures 8 and 9 illustrate the alignment of DNase hypersensitivity data with mouse-human conservation scores produced by AVID and visualized with

rVista across the ~90kb beta-globin locus and the T-cell receptor alpha LCR on chromosome 6.

5.8 Computer Readable Media and Programs Related to Regulatory sequence Profiles

Information regarding the identification, location, and/or activities of RSs may be stored and used in the form of computer readable medium. Furthermore, RS profiles may also be stored in computer readable medium. Accordingly, the invention provides computer readable medium comprising RSs and/or their genomic locations, and it provides computer readable medium comprising regulatory sequence profiles, including those generated by the methods of the invention. In certain embodiments, the computer readable medium comprises an regulatory sequence profile associated with a genetic locus, which may include an open reading frame. In one embodiment, the computer readable medium comprises the identification of known genes and their corresponding RS profiles in a particular cell. In another embodiment, the computer readable medium comprises the identification of a known gene and its corresponding RS profile in one or a plurality of cells, *e.g.* different cell types, diseased and normal cells, or cells treated with different agents. In certain preferred embodiments, the known genes are associated with a particular disease or disorder, such as a cancer for example. Specific known genes that may be included, include, but are not limited to, p53, Rb, INK4A/p16, CTNNB1, H-Ras, Fos, MDM2, INK4, ARF1, PTEN, Jun, WNT3A/14, NFkB, TERT, BRCA1, BRCA2, WAF1/p21, CDK4, TGF-beta1, RAR, E2F, VHL, MLH1, SMAD4, SMAD2, SMAD3, K-Ras, EGFR, WT1, Myc, Raf, ABL, and HER2. RS profiles may include numerical values corresponding to the activity of RSs within the profiled genetic locus.

Many of the methods of the invention are amenable to being performed by a computer program. Accordingly, the invention provides computer executable programs for accompanying one or more steps of any method of the invention. For methods related to comparing one or more RS profiles, the invention provides a computer executable program for comparing RS profiles, comprising inputting at least two RS profiles; comparing the values associated with each, and outputting a comparison of the two or more profiles. In a related embodiment, the invention provides a computer executable program for comparing an RS profile to one or more RS profiles located on computer readable media, comprising inputting the RS profile,

comparing the values of the RS profile to those of one or more profiles stored on a data set or computer readable media, and outputting a comparison of the comparison. In certain embodiment, the program may further identify RSs having different activities between compared RS profiles. Typically, comparisons will be performed between profiles established for one or more of the same genetic loci.

In another embodiment, the invention includes a computer executable program for profiling a genetic locus for active chromatin, comprising inputting data comprising regions of chromatin hypersensitivity sites derived from a selected cell or tissue type; comparing said data with data derived from the different cell or tissue type or with a control data set; and outputting at least one sequence associated with said locus or a genomic location of said active chromatin. In various embodiments, the inputted data may comprise sequences of chromatin hypersensitive sites generated by enzymatic digestion of chromatin or chromatin hypersensitive sites generated by using thermostable polymerase amplification of preselected regions of the genome. The preselected regions may be within 1, 5, 10, 25, 50, 100, or 200 kb of a gene known to be associated with a disease state.

In a related embodiment, the invention includes a computer executable program for profiling a genetic locus for allelic variants affecting the formation of active chromatin, comprising inputting data comprising regions of chromatin hypersensitivity sites derived from a selected mammalian cell or tissue type; comparing said data with data derived from the same cell or tissue type isolated from another mammal of the same species with a control data set representing normal or expected sequences from said species; and outputting at least one sequence having an allelic variant affecting said active chromatin formation.

5.9 Polynucleotides

Polynucleotides of the invention include polynucleotides comprising at least a portion of or a full length RS, regulatory sequence, regulatory unit, or variant thereof. The terms "DNA" and "polynucleotide" are used essentially interchangeably herein to refer to a DNA molecule that has been isolated free of total genomic DNA of a particular species. "Isolated," as used herein, refers to a polynucleotide that is substantially purified from other coding sequences, and that the DNA molecule does not contain large portions of unrelated coding DNA, such as large chromatin fragments or other functional genes or polypeptide coding regions. Of course, this

refers to the DNA molecule as originally isolated, and does not exclude genes or coding regions later added to the segment by the hand of man.

As will be understood by those skilled in the art, the polynucleotide compositions of this invention can include genomic sequences, extra-genomic and plasmid-encoded sequences and smaller engineered gene segments that express, or may be adapted to express, proteins, polypeptides, peptides and the like. Such segments may be naturally isolated, or modified synthetically by the hand of man.

As will be also recognized by the skilled artisan, polynucleotides of the invention may be single-stranded (coding or antisense) or double-stranded, and may be DNA (genomic, cDNA or synthetic) or RNA molecules. RNA molecules may include, for example, double-stranded RNA molecules, HnRNA molecules, which contain introns and correspond to a DNA molecule in a one-to-one manner, and mRNA molecules, which usually do not contain introns. Additional coding or non-coding sequences may, but need not, be present within a polynucleotide of the present invention, and a polynucleotide may, but need not, be linked to other molecules and/or support materials.

"Regulatory units" are polynucleotides that comprise sequences governing the expression of any given gene. Regulatory units may include one or more than one identifiable regulatory sequences. Regulatory units may be described functionally to include one, a plurality of, or all of the sequences involved in regulating the expression of a specific gene. Regulatory units include both polynucleotides comprising contiguous stretches of genomic nucleic acid sequence comprising one or more regulatory sequences associated with a gene and polynucleotides comprising one or more nucleic acid sequences that are not contiguous in the genome, wherein each of the sequences comprises one or more regulatory sequences associated with the same gene. The regulatory sequences of a regulatory unit may function cooperatively to affect transcription. For example, one or more regulatory sequences may coordinately increase transcription by recruiting a transcription factor and/or a polymerase. Alternatively, regulatory sequences within a regulatory unit may have different affects upon expression of their regulated gene. For example, one regulatory sequence may become active in response to an external stimuli and increase gene expression, while another regulatory sequence may be active in response to the same or a different stimuli to decrease transcription. In certain situations, one regulatory sequence may act to increase gene expression, while a second regulatory sequence

acts to decrease gene expression. Such apparently opposing functions are understood to play important roles in fine-tuning gene expression.

Regulatory units may also be described positionally to include all identified regulatory sequences within a certain distance in the genome from a specific gene. In certain embodiments, the regulatory sequences of a regulatory unit may be within 100 base pairs of a specific gene, within 500 base pairs of a specific gene, within 1000 base pairs of a specific gene, within 5000 base pairs of a specific gene, within 10,000 base pairs of a specific gene, within 50,000 base pairs of a specific gene, within 100,000 base pairs of a specific gene, or within 500,000 base pairs of a specific gene, for example. In certain embodiments, regulatory units comprise two or more regulatory sequences with similar or the same chromatin structure, DNase I or chemical hypersensitivity, or association with the same polypeptides, such as, for example, transcriptional activators, repressors, coactivators and corepressors. A variety of polypeptides had been identified that are involved in regulating transcription through association, either direct or indirect, with regulatory sequences, including, for example, the coactivator, CBP, and the corepressor, Sin3.

"Regulatory sequences" are nucleic acid sequences that are capable, alone or in combination, of affecting the expression of an associated gene; the pattern of expression between or among tissues; the timing of expression during development and differentiation; and the regulation of expression in response to external stimuli such as endogenous signaling molecules, or exogenous molecules including environmental and pharmaceutical agents, compounds, and chemicals. The gene may be regulated by the regulatory sequence in its natural position in the genome, or it may be regulated by the sequence in an artificial polynucleotide, such as an expression vector, for example. Regulatory sequences may mediate an increase or decrease in gene expression, and many regulatory sequences are capable of mediating either an increase or decrease in gene expression, depending upon the state of the cell. For example, a sequence first identified as the binding site for the Myc transcriptional activator, CACGTG, is capable of binding both transcriptional activators, including Myc proteins, and transcriptional repressors, including Mad family members. Typically, Myc proteins are expressed during cell growth, while Mad proteins are expressed during cell differentiation. Regulatory sequences include both promoter and enhancer sequences. Promoter sequences typically are associated with polymerase recruitment and are position-dependent in their ability to function.

Enhancer sequences typically bind one or more transcription factors (activators or repressors), which, in turn, activate or repress gene expression. Enhancer sequences frequently can function at a variety of positions relative to and distances from a regulated gene.

The invention includes polynucleotides consisting of or comprising functional fragments of regulatory sequences. It is understood that functional regulatory sequences may be very short, *e.g.* approximately six base pairs in length and that identified RSs or regulatory sequences may include shorter sequences that may function independently as a regulatory sequence. Functional fragments of the invention may, therefore, be of any length, including, for example, 5-10 nucleotides or base pairs, 10-20 nucleotides or base pairs, 20-50 nucleotides or base pairs, 50-200 nucleotides or base pairs, or greater than 200 nucleotides or base pairs, including any integer value between, such as 6 nucleotides or base pairs, for example. Regulatory sequences and RSs may include both core regulatory sequences and flanking sequence, which may or may not contribute to gene regulation. Accordingly, fragments of regulatory sequences or RSs that retain some or all functional activity, are included in the invention. Functional activity may be defined by any of a variety of different ways, including, for example, the ability to regulate expression of an associated gene, the ability to bind a transcription factor, the ability to direct chromatin structure, or the ability to recruit a coactivator, corepressor, or polymerase. The functional fragment may have the same activity as the larger regulatory sequence from which it was derived, or it may have less or greater activity. Typically, the functional fragment will have at least 25% to 1000% of the activity of the identified regulatory sequence when coupled to a reporter gene in the same manner.

Functional fragments may be identified by any means known and available in the art, including, for example, by sequence identification, functionally, or biochemical properties. For example, certain enhancer sequences are inverted repeats, so a functional fragment may be identified as an inverted repeat sequence within a regulatory sequence. Functional fragments may also be identified based upon their ability to direct transcription of an associated reporter gene. For example, discrete fragments of an identified regulatory sequence may be coupled to a reporter and tested for their affect on reporter expression. Alternatively, different site-specific mutations may be made throughout the regulatory sequence and their affect on the sequence's ability to direct expression of an associated reporter determined and used

to identify necessary sequences corresponding to a functional fragment. In addition, functional fragments may be identified based upon their ability to bind a polypeptide, using routine techniques such as electrophoretic mobility shift and footprinting assays.

In additional embodiments, the present invention provides polynucleotide fragments comprising various lengths of contiguous stretches of sequence identical to or complementary to one or more of the sequences disclosed herein. For example, polynucleotides are provided by this invention that comprise at least about 10, 15, 20, 30, 40, 50, 75, 100, 150, 200, 300, 400, 500 or 1000 or more contiguous nucleotides of one or more of the sequences disclosed herein as well as all intermediate lengths there between. It will be readily understood that "intermediate lengths", in this context, means any length between the quoted values, such as 16, 17, 18, 19, *etc.*; 21, 22, 23, *etc.*; 30, 31, 32, *etc.*; 50, 51, 52, 53, *etc.*; 100, 101, 102, 103, *etc.*; 150, 151, 152, 153, *etc.*; including all integers through 200-500; 500-1,000, and the like.

The invention also includes variants of regulatory sequences or RSs. Particularly in light of the discussion above, indicating that only a fragment of an identified regulatory sequence may be necessary for functional activity, it is understood that the identified regulatory sequences may be significantly altered, while retaining their ability to regulate gene expression. Furthermore, it has been demonstrated that even core enhancer elements may still bind transcription factors and regulate gene expression when one or more nucleotides is altered. Such alterations may include the deletion, insertion, or substitution of one or more nucleotides. Thus, the invention further includes variants of functional fragments of identified regulatory sequences. It is understood that one skilled in the art could readily determine what alterations could be made to a regulatory sequence or unit of the invention using routine procedures, such as mutagenesis, for example, and that screening for functional variants would not require undue experimentation, particularly given the relatively small size of regulatory sequences, and their core functional regions.

In certain embodiments, the present invention provides polynucleotide variants having substantial identity to sets of genome locations of regulatory sequences, including, for example, those comprising at least 70% sequence identity, preferably at least 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% or higher, wherein sequence identity is compared to a polynucleotide sequence of this invention

using the methods described herein, (*e.g.*, BLAST analysis using standard parameters, as described below).

Typically, polynucleotide variants will contain one or more substitutions, additions, deletions and/or insertions, preferably such that the functional activity of the variant polynucleotide is not substantially diminished relative to a polynucleotide sequence specifically set forth herein. The term "variants" should also be understood to encompass homologous sequences of xenogenic origin.

In another embodiment of the invention, polynucleotide compositions are provided that are capable of hybridizing under moderate to high stringency conditions to a polynucleotide sequence provided herein, or a fragment thereof, or a complementary sequence thereof. Hybridization techniques are well known in the art of molecular biology. For purposes of illustration, suitable moderately stringent conditions for testing the hybridization of a polynucleotide of this invention with other polynucleotides include prewashing in a solution of 5 X SSC, 0.5% SDS, 1.0 mM EDTA (pH 8.0); hybridizing at 50°C-60°C, 5 X SSC, overnight; followed by washing twice at 65°C for 20 minutes with each of 2X, 0.5X and 0.2X SSC containing 0.1% SDS. One skilled in the art will understand that the stringency of hybridization can be readily manipulated, such as by altering the salt content of the hybridization solution and/or the temperature at which the hybridization is performed. For example, in another embodiment, suitable highly stringent hybridization conditions include those described above, with the exception that the temperature of hybridization is increased, *e.g.*, to 60-65°C or 65-70°C. In certain embodiments, a positive hybridization is at least twice background. Those of ordinary skill will readily recognize that alternative hybridization and wash conditions can be utilized to provide conditions of similar stringency.

The polynucleotides of the present invention, or fragments thereof, regardless of the length of the coding sequence itself, may be combined with other DNA sequences, such as promoters, polyadenylation signals, additional restriction enzyme sites, multiple cloning sites, other coding segments, and the like, such that their overall length may vary considerably. It is therefore contemplated that a nucleic acid fragment of almost any length may be employed, with the total length preferably being limited by the ease of preparation and use in the intended recombinant DNA protocol. For example, illustrative polynucleotide segments with total lengths of about 10,000, about 5000, about 3000, about 2,000, about 1,000, about 500, about

200, about 100, about 50 base pairs in length, and the like, (including all intermediate lengths) are contemplated to be useful in many implementations of this invention.

When comparing polynucleotide sequences, two sequences are said to be "identical" if the sequence of nucleotides in the two sequences is the same when aligned for maximum correspondence, as described below. Comparisons between two sequences are typically performed by comparing the sequences over a comparison window to identify and compare local regions of sequence similarity. A "comparison window" as used herein, refers to a segment of at least about 20 contiguous positions, usually 30 to about 75, 40 to about 50, in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned.

Optimal alignment of sequences for comparison may be conducted using the Megalign program in the Lasergene suite of bioinformatics software (DNASTAR, Inc., Madison, WI), using default parameters. This program embodies several alignment schemes described in the following references: Dayhoff, M.O. (1978) A model of evolutionary change in proteins – Matrices for detecting distant relationships. In Dayhoff, M.O. (ed.) Atlas of Protein Sequence and Structure, National Biomedical Research Foundation, Washington DC Vol. 5, Suppl. 3, pp. 345-358; Hein J. (1990) Unified Approach to Alignment and Phylogenes pp. 626-645 *Methods in Enzymology* vol. 183, Academic Press, Inc., San Diego, CA; Higgins, D.G. and Sharp, P.M. (1989) *CABIOS 5*:151-153; Myers, E.W. and Muller W. (1988) *CABIOS 4*:11-17; Robinson, E.D. (1971) *Comb. Theor 11*:105; Santou, N. Nes, M. (1987) *Mol. Biol. Evol. 4*:406-425; Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy – the Principles and Practice of Numerical Taxonomy*, Freeman Press, San Francisco, CA; Wilbur, W.J. and Lipman, D.J. (1983) *Proc. Natl. Acad., Sci. USA 80*:726-730.

Alternatively, optimal alignment of sequences for comparison may be conducted by the local identity algorithm of Smith and Waterman (1981) *Add. APL. Math* 2:482, by the identity alignment algorithm of Needleman and Wunsch (1970) *J. Mol. Biol.* 48:443, by the search for similarity methods of Pearson and Lipman (1988) *Proc. Natl. Acad. Sci. USA* 85: 2444, by computerized implementations of these algorithms (GAP, BESTFIT, BLAST, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, WI), or by inspection.

One preferred example of algorithms that are suitable for determining percent sequence identity and sequence similarity are the BLAST and BLAST 2.0 algorithms, which are described in Altschul et al. (1977) *Nucl. Acids Res.* 25:3389-3402 and Altschul et al. (1990) *J. Mol. Biol.* 215:403-410, respectively. BLAST and BLAST 2.0 can be used, for example with the parameters described herein, to determine percent sequence identity for the polynucleotides of the invention. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information. In one illustrative example, cumulative scores can be calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0) and N (penalty score for mismatching residues; always <0). Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, and expectation (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff and Henikoff (1989) *Proc. Natl. Acad. Sci. USA* 89:10915) alignments, (B) of 50, expectation (E) of 10, M=5, N=-4 and a comparison of both strands.

Preferably, the "percentage of sequence identity" is determined by comparing two optimally aligned sequences over a window of comparison of at least 20 positions, wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (*i.e.,* gaps) of 20 percent or less, usually 5 to 15 percent, or 10 to 12 percent, as compared to the reference sequences (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid bases occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the reference sequence (*i.e.,* the window size) and multiplying the results by 100 to yield the percentage of sequence identity.

Further, alleles of the genes comprising the polynucleotide sequences provided herein are within the scope of the present invention. Alleles are endogenous genes that may be altered as a result of one or more mutations, such as deletions, additions and/or substitutions of nucleotides. The resulting mRNA and protein may, but need

not, have an altered structure or function. Alleles may be identified using standard techniques such as hybridization, amplification and/or database sequence comparison.

The loci information presented in regulatory sequence profiles reveals not only regulatory sequences and RSs, but also shows how the sequences, and fragments and variants thereof, may be used for regulation in the genome. For example, a regulatory sequence that is immediately upstream and contiguous to a regulated gene regulates a gene when the gene is immediately downstream and contiguous to the regulatory sequence. A regulatory sequence that is downstream and separated from its matching open reading frame by 200 bases may be used to regulate a gene by placement downstream and separated from the gene by about 200 bases. In each instance, a regulatory sequence may be used to regulate genes other than the ones that are matched in the table, by respecting the same distance, orientation and placement criteria as shown by the locus information for each row in the table. The position of the regulatory sequence with respect to the regulated gene and/or the distance of the regulatory sequence from the regulated gene may be identical to the position or distance as compared to the genomic position of the regulatory sequence and an associated gene. However, it is also understood that the position and/or distance may be varied significantly and still retain the ability to regulate gene expression. Thus, a regulatory sequence may be located at a similar or different position or distance from a regulated gene. For example, a regulatory element may be located either upstream or downstream of a regulated gene, and it may be located nearer to or further from a regulated gene, so long as it retains the ability to regulate expression of an associated gene. In certain embodiments, a different position is within 100 base pairs, within 500 base pairs, within 1000 base pairs, within 3000 base pairs, or greater than 3000 base pairs, and all intermediate lengths, as compared to the endogenous or genomic distance from a regulated gene. The ability of a regulatory element to regulate expression of an associated gene may be determined by routine procedures, including, for example, placing the regulatory sequence at different positions relative to a reporter gene in an expression construct and determining the effects of the position of the regulatory sequence on expression of the reporter.

In other embodiments of the present invention, the polynucleotide sequences provided herein can be advantageously used as probes or primers for nucleic acid hybridization. As such, it is contemplated that nucleic acid segments that comprise a sequence region of at least about 10 or 15 nucleotide long contiguous sequence that

has the same sequence as, or is complementary to, a 10 or 15 nucleotide long contiguous sequence disclosed herein will find particular utility. Longer contiguous identical or complementary sequences, *e.g.*, those of about 20, 30, 40, 50, 100, 200, 500, 1000 (including all intermediate lengths) and even up to full-length sequences will also be of use in certain embodiments. In certain embodiments, oligonucleotides may comprise at least 10 bases, 10-75 bases or 12-30 bases.

The ability of such nucleic acid probes to specifically hybridize to a sequence of interest will enable them to be of use in detecting the presence of complementary sequences in a given sample. However, other uses are also envisioned, such as the use of the sequence information for the preparation of mutant species primers, or primers for use in preparing other genetic constructs and polynucleotides.

Polynucleotide molecules having sequence regions consisting of contiguous nucleotide stretches of 5-9, 10-14, 15-20, 30, 50, or even of 100-200 nucleotides or so (including intermediate lengths as well), identical, substantially complementary or completely complementary to a polynucleotide sequence disclosed herein, are particularly contemplated, for example, as hybridization probes for use in, *e.g.*, Southern and Northern blotting. This would allow a gene product, or fragment thereof, to be analyzed, both in diverse cell types and also in various bacterial cells. The total size of fragment, as well as the size of the complementary stretch(es), will ultimately depend on the intended use or application of the particular nucleic acid segment. Smaller fragments will generally find use in hybridization embodiments, wherein the length of the contiguous complementary region may be varied, such as between about 15 and about 100 nucleotides, but larger contiguous complementarity stretches may be used, according to the length complementary sequences one wishes to detect.

The use of a hybridization probe of about 15-25 nucleotides in length allows the formation of a duplex molecule that is both stable and selective. Molecules having contiguous complementary sequences over stretches greater than 15 bases in length are generally preferred, though, in order to increase stability and selectivity of the hybrid, and thereby improve the quality and degree of specific hybrid molecules obtained. One will generally prefer to design nucleic acid molecules having gene-complementary stretches of 15 to 25 contiguous nucleotides, or even longer where desired.

Hybridization probes may be selected from any portion of any of the sequences disclosed herein. All that is required is to review the sequences set forth herein, or to any continuous portion of the sequences, from about 15-25 nucleotides in length up to and including the full length sequence, that one wishes to utilize as a probe or primer. The choice of probe and primer sequences may be governed by various factors. For example, one may wish to employ primers from towards the termini of the total sequence.

Small polynucleotide segments or fragments may be readily prepared by, for example, directly synthesizing the fragment by chemical means, as is commonly practiced using an automated oligonucleotide synthesizer. Also, fragments may be obtained by application of nucleic acid reproduction technology, such as the PCR™ technology of U. S. Patent 4,683,202 (incorporated herein by reference), by introducing selected sequences into recombinant vectors for recombinant production, and by other recombinant DNA techniques generally known to those of skill in the art of molecular biology.

The nucleotide sequences of the invention may be used for their ability to selectively form duplex molecules with complementary stretches of the entire gene or gene fragments of interest. Depending on the application envisioned, one will typically desire to employ varying conditions of hybridization to achieve varying degrees of selectivity of probe towards target sequence. For applications requiring high selectivity, one will typically desire to employ relatively stringent conditions to form the hybrids, *e.g.*, one will select relatively low salt and/or high temperature conditions, such as provided by a salt concentration of from about 0.02 M to about 0.15 M salt at temperatures of from about 50°C to about 70°C. Such selective conditions tolerate little, if any, mismatch between the probe and the template or target strand, and would be particularly suitable for isolating related sequences.

Of course, for some applications, for example, where one desires to prepare mutants employing a mutant primer strand hybridized to an underlying template, less stringent (reduced stringency) hybridization conditions will typically be needed in order to allow formation of the heteroduplex. In these circumstances, one may desire to employ salt conditions such as those of from about 0.15 M to about 0.9 M salt, at temperatures ranging from about 20°C to about 55°C. Cross-hybridizing species can thereby be readily identified as positively hybridizing signals with respect to control

hybridizations. In any case, it is generally appreciated that conditions can be rendered more stringent by the addition of increasing amounts of formamide, which serves to destabilize the hybrid duplex in the same manner as increased temperature. Thus, hybridization conditions can be readily manipulated, and thus will generally be a method of choice depending on the desired results.

Treatment designed to reduce the levels of expression of a gene, including those identified as described above, may also act by reducing expression, for example, using knockout or knockdown reagents. The invention provides knockdown and knockout reagent, as well as knockout cells, plants, and animals produced using such reagents. Knockout reagents include targeting or homologous recombination vectors specific for at least a region of the gene being targeted. In certain embodiments, the region is a hypersensitivity site or active control region of the invention. The invention also includes transgenic and knockout cells, plants, and animals comprising a disrupted nucleic acid sequence of the invention. Transgenics and knockouts of the invention include any suitable plant or animal, including humans and other mammals, such as mice, for example. In one embodiment, the invention includes a transgenic animal that expresses a polynucleic acid or polypeptide, wherein expression is regulated by a regulatory element or unit of the invention. In another embodiment, one or more regulatory elements or units are disrupted in a cell or an animal using knockout methods, such that expression of a gene regulated by the disrupted sequence(s) is altered. Methods for obtaining transgenic and knockout cells and animals are well known in the art. Methods of generating transgenic animals are described, for example, in Hofker, M.H. (ed.), Van Deursen, J., and Sklar, H.T., (2002), TRANSGENIC MOUSE: METHODS AND PROTOCOLS (METHODS IN MOLECULAR BIOLOGY), Humana Press, Clifton, NJ. Methods of generating a mouse containing an introduced gene disruption are described, for example, in Hogan, B. *et al.*, (1994), MANIPULATING THE MOUSE EMBRYO: A LABORATORY MANUAL, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. In one embodiment, gene targeting, which is a method of using homologous recombination to modify a cell's or animal's genome, can be used to introduce changes into cultured embryonic stem cells. By targeting a nucleic acid sequence of interest in ES cells, these changes can be introduced into the germlines of animals to generate chimeras and knockout animals. Transgenic cells and animals of the invention are particularly useful in providing or expressing a functional polypeptide in a particular cell or at a specific

time in development or cell cycle, for example. A nucleic acid of the invention may be chosen to direct gene expression based upon the identification of the cell types and times during which it is active or hypersensitive. Knockout cells and animals of the invention are useful in identifying genes regulated by the disrupted nucleic acid of the invention and the function of the disrupted nucleic acid of the invention.

Knockdown reagents include any of a variety of agents that may reduce mRNA levels. Knockdown reagents include, for example, ribozymes, antisense RNA, and double-stranded RNAs, including small interfering RNAs (siRNAs) and short hairpin RNAs (shRNAs).

Antisense oligonucleotides had been demonstrated to be effective and targeted inhibitors of protein synthesis, and, consequently, can be used to specifically inhibit protein synthesis by a targeted gene. The efficacy of antisense oligonucleotides for inhibiting protein synthesis is well established. For example, the synthesis of polygalactauronase and the muscarine type 2 acetylcholine receptor are inhibited by antisense oligonucleotides directed to their respective mRNA sequences (U. S. Patent 5,739,119 and U. S. Patent 5,759,829). Further, examples of antisense inhibition had been demonstrated with the nuclear protein cyclin, the multiple drug resistance gene (MDG1), ICAM-1, E-selectin, STK-1, striatal $GABA_A$ receptor and human EGF (Jaskulski *et al.*, Science. 1988 Jun 10;240(4858):1544-6; Vasanthakumar and Ahmed, Cancer Commun. 1989;1(4):225-32; Peris *et al.*, Brain Res Mol Brain Res. 1998 Jun 15;57(2):310-20; U. S. Patent 5,801,154; U.S. Patent 5,789,573; U. S. Patent 5,718,709 and U.S. Patent 5,610,288). Furthermore, antisense constructs have also been described that inhibit and can be used to treat a variety of abnormal cellular proliferations, *e.g.* cancer (U. S. Patent 5,747,470; U. S. Patent 5,591,317 and U. S. Patent 5,783,683).

Therefore, in certain embodiments, the present invention provides oligonucleotide sequences that comprise all, or a portion of, any sequence that is capable of specifically binding to a selected target polynucleotide sequence, or a complement thereof. In one embodiment, the antisense oligonucleotides comprise DNA or derivatives thereof. In another embodiment, the oligonucleotides comprise RNA or derivatives thereof. The antisense oligonucleotides may be modified DNAs comprising a phosphorothioated modified backbone. Also, the oligonucleotide sequences may comprise peptide nucleic acids or derivatives thereof. In each case, preferred compositions comprise a sequence region that is complementary, and more

preferably, completely complementary to one or more portions of a target gene or polynucleotide sequence. Selection of antisense compositions specific for a given sequence is based upon analysis of the chosen target sequence and determination of secondary structure, T_m, binding energy, and relative stability. Antisense compositions may be selected based upon their relative inability to form dimers, hairpins, or other secondary structures that would reduce or prohibit specific binding to the target mRNA in a host cell. Highly preferred target regions of the mRNA include those regions at or near the AUG translation initiation codon and those sequences that are substantially complementary to 5' regions of the mRNA. These secondary structure analyses and target site selection considerations can be performed, for example, using v.4 of the OLIGO primer analysis software and/or the BLASTN 2.0.5 algorithm software (Altschul *et al.*, Nucleic Acids Res. 1997, 25(17):3389-402).

According to another embodiment of the invention, ribozyme molecules are used to inhibit expression of a target gene or polynucleotide sequence. Ribozymes are RNA-protein complexes that cleave nucleic acids in a site-specific fashion. Ribozymes have specific catalytic domains that possess endonuclease activity (Kim and Cech, Proc Natl Acad Sci U S A. 1987 Dec;84(24):8788-92; Forster and Symons, Cell. 1987 Apr 24;49(2):211-20). For example, a large number of ribozymes accelerate phosphoester transfer reactions with a high degree of specificity, often cleaving only one of several phosphoesters in an oligonucleotide substrate (Cech *et al.*, Cell. 1981 Dec;27(3 Pt 2):487-96; Michel and Westhof, J Mol Biol. 1990 Dec 5;216(3):585-610; Reinhold-Hurek and Shub, Nature. 1992 May 14;357(650):173-6). This specificity has been attributed to the requirement that the substrate bind via specific base-pairing interactions to the internal guide sequence ("IGS") of the ribozyme prior to chemical reaction.

At least six basic varieties of naturally-occurring enzymatic RNAs are known presently. Each can catalyze the hydrolysis of RNA phosphodiester bonds *in trans* (and thus can cleave other RNA molecules) under physiological conditions. In general, enzymatic nucleic acids act by first binding to a target RNA. Such binding occurs through the target binding portion of an enzymatic nucleic acid that is held in close proximity to an enzymatic portion of the molecule that acts to cleave the target RNA. Thus, the enzymatic nucleic acid first recognizes and then binds a target RNA through complementary base-pairing, and once bound to the correct site, acts

enzymatically to cut the target RNA. Strategic cleavage of such a target RNA will destroy its ability to direct synthesis of an encoded protein. After an enzymatic nucleic acid has bound and cleaved its RNA target, it is released from that RNA to search for another target and can repeatedly bind and cleave new targets.

The enzymatic nature of a ribozyme may be advantageous over many technologies, such as antisense technology (where a nucleic acid molecule simply binds to a nucleic acid target to block its translation), since the concentration of ribozyme necessary to affect inhibition of expression is lower than that of an antisense oligonucleotide. This advantage reflects the ability of the ribozyme to act enzymatically. Thus, a single ribozyme molecule is able to cleave many molecules of target RNA. In addition, the ribozyme is a highly specific inhibitor, with the specificity of inhibition depending not only on the base pairing mechanism of binding to the target RNA, but also on the mechanism of target RNA cleavage. Single mismatches, or base-substitutions, near the site of cleavage can completely eliminate catalytic activity of a ribozyme. Similar mismatches in antisense molecules do not prevent their action (Woolf *et al.*, 1992, *Proc. Natl. Acad. Sci. U S A.* 89: 7305-9). Thus, the specificity of action of a ribozyme is greater than that of an antisense oligonucleotide binding the same RNA site.

The enzymatic nucleic acid molecule may be formed in a hammerhead, hairpin, a hepatitis δ virus, group I intron or RNaseP RNA (in association with an RNA guide sequence) or Neurospora VS RNA motif, for example. Specific examples of hammerhead motifs are described by Rossi *et al.*, 1992, *Nucleic Acids Res.* 20: 4559-65. Examples of hairpin motifs are described by Hampel *et al.* (Eur. Pat. Appl. Publ. No. EP 0360257), Hampel and Tritz, Biochemistry 1989 Jun 13;28(12):4929-33; Hampel *et al.*, Nucleic Acids Res. 1990 Jan 25;18(2):299-304 and U. S. Patent 5,631,359. An example of the hepatitis δ virus motif is described by Perrotta and Been, Biochemistry. 1992 Dec 1;31(47):11843-52; an example of the RNaseP motif is described by Guerrier-Takada *et al.*, Cell. 1983 Dec;35(3 Pt 2):849-57; Neurospora VS RNA ribozyme motif is described by Collins (Saville and Collins, Cell. 1990 May 18;61(4):685-96; Saville and Collins, Proc Natl Acad Sci U S A. 1991 Oct 1;88(19):8826-30; Collins and Olive, Biochemistry. 1993 Mar 23;32(11):2795-9); and an example of the Group I intron is described in (U. S. Patent 4,987,071). Important characteristics of enzymatic nucleic acid molecules used according to the invention

are that they have a specific substrate binding site which is complementary to one or more of the target gene DNA or RNA regions, and that they have nucleotide sequences within or surrounding that substrate binding site which impart an RNA cleaving activity to the molecule. Thus the ribozyme constructs need not be limited to specific motifs mentioned herein.

Ribozymes may be designed as described in Int. Pat. Appl. Publ. No. WO 93/23569 and Int. Pat. Appl. Publ. No. WO 94/02595, each specifically incorporated herein by reference, and synthesized to be tested *in vitro* and *in vivo*, as described. Such ribozymes can also be optimized for delivery. While specific examples are provided, those in the art will recognize that equivalent RNA targets in other species can be utilized when necessary.

Ribozyme activity can be optimized by altering the length of the ribozyme binding arms, or chemically synthesizing ribozymes with modifications that prevent their degradation by serum ribonucleases (see *e.g.*, Int. Pat. Appl. Publ. No. WO 92/07065; Int. Pat. Appl. Publ. No. WO 93/15187; Int. Pat. Appl. Publ. No. WO 91/03162; Eur. Pat. Appl. Publ. No. 92110298.4; U. S. Patent 5,334,711; and Int. Pat. Appl. Publ. No. WO 94/13688, which describe various chemical modifications that can be made to the sugar moieties of enzymatic RNA molecules), modifications which enhance their efficacy in cells, and removal of stem II bases to shorten RNA synthesis times and reduce chemical requirements.

RNA interference methods using double-stranded RNA also may be used to disrupt the expression of a gene or polynucleotide of interest. A dsRNA molecule that targets and induces degradation of an mRNA that is derived from a gene or polynucleotide of interest can be introduced into a cell. The exact mechanism of how the dsRNA targets the mRNA is not essential to the operation of the invention, other than the dsRNA shares sequence homology with the mRNA transcript. The mechanism could be a direct interaction with the target gene, an interaction with the resulting mRNA transcript, an interaction with the resulting protein product, or another mechanism. Again, while the exact mechanism is not essential to the invention, it is believed the association of the dsRNA to the target gene is defined by the homology between the dsRNA and the actual and/or predicted mRNA transcript. It is believed that this association will affect the ability of the dsRNA to disrupt the target gene. dsRNA methods and reagents are described in PCT application WO

01/68836, WO 01/29058, WO 02/44321, and WO 01/75164, which are hereby incorporated by reference in their entirety.

In one embodiment of the invention, double-stranded RNA interference (dsRNAi) may be used to specifically inhibit target nucleic acid expression. Briefly, it is hypothesized that the presence of double-stranded RNA dominantly silences gene expression in a sequence-specific manner by causing the corresponding RNA to be degraded. Although first discovered in lower organisms such as the nematode and *Drosphila*, for example, dsRNAi has also been demonstrated to work in fungi, plants, and mammalian cells (Wianny, F. and Zernica-Goetz, M. (2000), Nature Cell Biology Vol. 2, 70-75). However, transfection of long dsRNAs into mammalian cells can result in nonspecific gene suppression, as opposed to the gene-specific suppression observed in other organisms.

Although the mechanisms behind dsRNAi is still not entirely understood, experiments demonstrated that, in the cell, a double-stranded RNA (dsRNA) is cleaved into short pieces, typically 21-25 nucleotides in length, termed small interfering RNAs (siRNAs), by a ribonuclease such as DICER. The siRNAs subsequently assemble with protein components into an RNA-induced silencing complex (RISC), which binds to and tags the complementary portion of the target mRNA for nuclease digestion. The siRNA triggers the degradation of mRNA that matches its sequence, thereby repressing expression of the corresponding gene. Discussed in Bass, B. Nature 411:428-429 (2001) and Sharp, P.A. Genes Dev. 15:485-490 (2001).

Double-stranded RNA-mediated suppression of gene and nucleic acid expression may be accomplished according to the invention by introducing dsRNA, siRNA or shRNA into cells or organisms. dsRNAs less than 30 nucleotides in length do not appear to induce nonspecific gene suppression, as described above for long dsRNA molecules. Indeed, the direct introduction of siRNAs to a cell can trigger RNAi in mammalian cells (Elshabir, S.M., *et al.* Nature 411:494-498 (2001)). Furthermore, suppression in mammalian cells occurred at the RNA level and was specific for the targeted genes, with a strong correlation between RNA and protein suppression (Caplen, N. *et al.*, Proc. Natl. Acad. Sci. USA 98:9746-9747 (2001)). In addition, it was shown that a wide variety of cell lines, including HeLa S3, COS7, 293, NIH/3T3, A549, HT-29, CHO-KI and MCF-7 cells, are susceptible to some level

of siRNA silencing (Brown, D. *et al.* TechNotes 9(1):1-7, available at http://www.ambion.com/techlib/tn/91/912.html (9/1/02)).

Structural characteristics of effective siRNA molecules had been identified. Elshabir, S.M. *et al.* (2001) Nature 411:494-498 and Elshabir, S.M. *et al.* (2001), EMBO 20:6877-6888. Accordingly, one of skill in the art would understand that a wide variety of different siRNA molecules may be used to target a specific gene or transcript. In certain embodiments, siRNA molecules according to the invention are 18 - 25 nucleotides in length, including each integer in between. In one embodiment, an siRNA is 21 nucleotides in length. In certain embodiments, siRNAs have 0-7 nucleotide 3' overhangs or 0-4 nucleotide 5' overhangs. In one embodiment, an siRNA molecule has a two nucleotide 3' overhang. In one embodiment, an siRNA is 21 nucleotides in length with two nucleotide 3' overhangs (*i.e.* they contain a 19 nucleotide complementary region between the sense and antisense strands). In certain embodiments, the overhangs are UU or dTdT 3' overhangs. Generally, siRNA molecules are completely complementary to one strand of a target DNA molecule, since even single base pair mismatches had been shown to reduce silencing. In other embodiments, siRNAs may have a modified backbone composition, such as, for example, 2'-deoxy- or 2'-O-methyl modifications. However, in preferred embodiments, the entire strand of the siRNA is not made with either 2' deoxy or 2'-O-modified bases.

Short hairpin RNAs may also be used to inhibit or knockdown gene or nucleic acid expression according to the invention. Short Hairpin RNA (shRNA) is a form of hairpin RNA capable of sequence-specifically reducing expression of a target gene. Short hairpin RNAs may offer an advantage over siRNAs in suppressing gene expression, as they are generally more stable and less susceptible to degradation in the cellular environment. It has been established that such short hairpin RNA-mediated gene silencing (also termed SHAGging) works in a variety of normal and cancer cell lines, and in mammalian cells, including mouse and human cells. Paddison, P. *et al.*, Genes Dev. 16(8):948-58 (2002). Furthermore, transgenic cell lines bearing chromatin genes that code for engineered shRNAs had been generated. These cells are able to constitutively synthesize shRNAs, thereby facilitating long-lasting or constitutive gene silencing that may be passed on to progeny cells. Paddison, P. *et al.*, Proc. Natl. Acad. Sci. USA 99(3):1443-1448 (2002).

ShRNAs contain a stem loop structure. In certain embodiments, they may contain variable stem lengths, typically from 19 to 29 nucleotides in length, or any number in between. In certain embodiments, hairpins contain 19 to 21 nucleotide stems, while in other embodiments, hairpins contain 27 to 29 nucleotide stems. In certain embodiments, loop size is between 4 to 23 nucleotides in length, although the loop size may be larger than 23 nucleotides without significantly affecting silencing activity. ShRNA molecules may contain mismatches, for example G-U mismatches between the two strands of the shRNA stem without decreasing potency. In fact, in certain embodiments, shRNAs are designed to include one or several G-U pairings in the hairpin stem to stabilize hairpins during propagation in bacteria, for example. However, complementarity between the portion of the stem that binds to the target mRNA (antisense strand) and the mRNA is typically required, and even a single base pair mismatch is this region may abolish silencing. 5' and 3' overhangs are not required, since they do not appear to be critical for shRNA function, although they may be present (Paddison *et al.* (2002) *Genes & Dev.* 16(8):948-58).

Since the invention provides regulatory units comprising one or more identified regulatory sequences, the invention further includes polynucleotides comprising regulatory units. In certain embodiments, these polynucleotides contain two or more regulatory sequences identified as belonging to the same regulatory unit or as regulating the same gene. In certain embodiments, they may comprise three, four or more regulatory sequences. Polynucleotides comprising regulatory units may include regulatory sequences positioned relative to each other or a gene at the same or a similar location as compared to their genomic relationship. The same position is the same number of bases from each other or a gene or within 100 base pairs of their genomic distance from each other or a gene, while a similar location is between 100 and 10,000, 25,000, or 50,000 base pairs from each other or a gene, each including any integer value between. For example, the β-globin locus 4 sites are clustered within a stretch of about 20 kb, and the genes that they regulate are 6, 20, 27 and 50 kb away. In certain embodiments, polynucleotides comprising regulatory units have each included regulatory sequence in the same order, 5' to 3', as found genomically. Furthermore, in certain embodiments, polynucleotides of the present invention comprise two or more regulatory units, each normally associated with a different gene. Thus, the invention includes polynucleotides with any combination of

regulatory sequences or regulatory units, as it may be advantageous to include such combinations to direct gene expression as desired, for example, to a particular cell type at a particular stage of the cell cycle, in order to effectively provide a therapeutic molecule to a cell with a disease or disorder.

The polynucleotides derived from or containing hypersensitive sites, and fragments and variants thereof, are useful in the regulation of gene expression. Accordingly, the invention contemplates polynucleotides comprising an aforementioned polynucleotide and an open reading frame. In certain embodiments, the regulatory sequence or regulatory unit is operatively linked to the open reading frame, for example, such that the regulatory sequence or unit regulates expression of the open reading frame.

The open reading frame may be any polynucleotide sequence of interest, including sequences capable of expressing RNA or polypeptides. In one embodiment, open reading frame may be a therapeutic molecule, such as a therapeutic polypeptide or knockdown reagent. Therapeutic molecules may be provided to replace a polypeptide lacking in a patient suffering from a disease or disorder or to inhibit expression of a gene overexpressed or inappropriately expressed in a patient suffering from a disease or disorder, for example. In other embodiments, the open reading frame may be a reporter gene and may encode a reporter molecule. Representative examples of reporter genes and molecules include those listed in Table 1, as well lacZ, neoR, dhfr, alphaIV, and uidA genes.

Table 1

Protein	**Activity & Measurement**
CAT (chloramphenicol acetyltransferase)	Transfers radioactive acetyl groups to chloramphenicol; detection by thin layer chromatography and autoradiography
GAL (β galactosidase)	Hydrolyzes colorless galactosides to yield colored products.
GUS (β glucuronidase)	Hydrolyzes colorless glucuronides to yield coloured products.
LUC (luciferase)	Oxidizes luciferin, emitting photons.
GFP (green fluorescent protein)	Fluoresces on irradiation with UV.

Reporter genes may be used to "report" many different properties and events, for example: (i) the strength of promoters, whether native or modified for reverse genetics studies; (ii) the efficiency of gene delivery systems; (iii) the intracellular fate of a gene product, a result of protein traffic; (iv) the interaction of two proteins in the

two-hybrid system or of a protein and a nucleic acid in the one-hybrid system; (v) the efficiency of translation initiation signals; and (vi) the success of molecular cloning efforts.

The invention also includes vectors and host cells comprising one or more polynucleotides of the invention. All types of vectors are included, including, but not limited to, expression vectors, gene trap vectors, homologous recombination or targeting vectors, and cloning vectors. Such transcription units can be incorporated into a variety of vectors for introduction into mammalian cells, including but not restricted to, plasmid DNA vectors, viral DNA vectors (such as adenovirus or adeno-associated vectors), or viral RNA vectors (such as retroviral, semliki forest virus, sindbis virus vectors). Vectors of the invention include, in certain embodiments, two or more regulatory sequences or RSs, including, for example, two or more RSs associated with a regulatory profile. In certain embodiments, the regulatory sequences or regulatory units include sequences found clustered or within a specified distance from each other in the genome. In another embodiment, the regulatory sequences or regulatory units include sequences that coordinately regulate gene expression, for example, in the genome or in a synthesize polynucleotide construct.

Regulatory elements and regulatory units of the invention may be used to drive expression of or produce a polypeptide. In order to express a desired polypeptide, the nucleotide sequences encoding the polypeptide, or functional equivalents, may be inserted into appropriate expression vector, *i.e.*, a vector that contains the necessary elements for the transcription and translation of the inserted coding sequence. Methods well known to those skilled in the art may be used to construct expression vectors containing sequences encoding a polypeptide of interest and appropriate transcriptional and translational control elements. These methods include *in vitro* recombinant DNA techniques, synthetic techniques, and *in vivo* genetic recombination. Such techniques are described, for example, in Sambrook, J. *et al.* (1989) *Molecular Cloning, A Laboratory Manual,* Cold Spring Harbor Press, Plainview, N.Y., and Ausubel, F. M. *et al.* (1989) *Current Protocols in Molecular Biology,* John Wiley & Sons, New York. N.Y.

A variety of expression vector/host systems may be utilized to contain and express polynucleotide sequences. These include, but are not limited to, microorganisms such as bacteria transformed with recombinant bacteriophage, plasmid, or cosmid DNA expression vectors; yeast transformed with yeast expression

vectors; insect cell systems infected with virus expression vectors (*e.g.*, baculovirus); plant cell systems transformed with virus expression vectors (*e.g.*, cauliflower mosaic virus, CaMV; tobacco mosaic virus, TMV) or with bacterial expression vectors (*e.g.*, Ti or pBR322 plasmids); or animal cell systems.

Vectors of the invention include one or more polynucleotides of the invention, which may be included to regulate gene expression. Such vectors may further comprise other regulatory elements involved in gene expression, including, for example, promoter and enhancer sequences, IRES sequences, and polyA sites. The "control elements" or "regulatory sequences" present in an expression vector may include a variety of non-translated regions of the vector--enhancers, promoters, 5' and 3' untranslated regions, which may interact with host cellular proteins to carry out transcription and translation. Such elements may vary in their strength and specificity. Depending on the vector system and host utilized, any number of suitable transcription and translation elements, including constitutive and inducible promoters, may be used. For example, when cloning in bacterial systems, inducible promoters such as the hybrid lacZ promoter of the PBLUESCRIPT phagemid (Stratagene, La Jolla, Calif.) or PSPORT1 plasmid (Gibco BRL, Gaithersburg, MD) and the like may be used.

In mammalian cell systems, promoters from mammalian genes or from mammalian viruses are generally preferred. If it is necessary to generate a cell line that contains multiple copies of the sequence encoding a polypeptide, vectors based on SV40 or EBV may be advantageously used with an appropriate selectable marker. In addition to expression vectors useful, for example, in the production of polypeptides, the invention also contemplates therapeutic agents, including gene therapy vectors, for example. In certain embodiments, gene therapy vectors of the invention include viral vectors and may be designed for either transient or stable expression of an encoded molecule.

The invention further includes gene trap and homologous recombination vectors designed for introducing a polynucleotide of the invention into a genome. Gene trap vectors may be use, for example to randomly insert a regulatory sequence or regulatory unit into the genome to alter expression of a gene. Collections or libraries of gene trapped cells may be used, for example, to screen for therapeutic targets and to screen drug candidates. Targeting vectors of the invention may be used, for example, to insert or replace a genomic sequence, for example, to alter the

expression of a gene. Such vectors may be used therapeutically, for example, to correct aberrant gene expression associated with a disease or disorder.

In bacterial systems, any of a number of expression vectors may be selected depending upon the use intended for the expressed polypeptide. For example, when large quantities are needed, for example for the induction of antibodies, vectors which direct high level expression of fusion proteins that are readily purified may be used. Such vectors include, but are not limited to, the multifunctional *E. coli* cloning and expression vectors such as BLUESCRIPT (Stratagene), in which the sequence encoding the polypeptide of interest may be ligated into the vector in frame with sequences for the amino-terminal Met and the subsequent 7 residues of □-galactosidase so that a hybrid protein is produced; pIN vectors (Van Heeke, G. and S. M. Schuster (1989) *J. Biol. Chem. 264*:5503-5509); and the like. pGEX Vectors (Promega, Madison, Wis.) may also be used to express foreign polypeptides as fusion proteins with glutathione S-transferase (GST). In general, such fusion proteins are soluble and can easily be purified from lysed cells by adsorption to glutathione-agarose beads followed by elution in the presence of free glutathione.

In the yeast, *Saccharomyces cerevisiae*, a number of vectors containing constitutive or inducible promoters such as alpha factor, alcohol oxidase, and PGH may be used. For reviews, see Ausubel et al. (supra) and Grant et al. (1987) *Methods Enzymol.* 153:516-544.

In cases where plant expression vectors are used, the expression of sequences encoding polypeptides may be driven by any of a number of promoters. For example, viral promoters such as the 35S and 19S promoters of CaMV may be used alone or in combination with the omega leader sequence from TMV (Takamatsu, N. (1987) *EMBO J. 6*:307-311. Alternatively, plant promoters such as the small subunit of RUBISCO or heat shock promoters may be used (Coruzzi, G. et al. (1984) *EMBO J. 3*:1671-1680; Broglie, R. et al. (1984) *Science 224*:838-843; and Winter, J. et al. (1991) *Results Probl. Cell Differ. 17*:85-105). These constructs can be introduced into plant cells by direct DNA transformation or pathogen-mediated transfection. Such techniques are described in a number of generally available reviews (see, for example, Hobbs, S. or Murry, L. E. in McGraw Hill Yearbook of Science and Technology (1992) McGraw Hill, New York, N.Y.; pp. 191-196).

An insect system may also be used to express a polypeptide of interest. For example, in one such system, Autographa californica nuclear polyhedrosis virus

(AcNPV) is used as a vector to express foreign genes in Spodoptera frugiperda cells or in Trichoplusia larvae. The sequences encoding the polypeptide may be cloned into a non-essential region of the virus, such as the polyhedrin gene, and placed under control of the polyhedrin promoter. Successful insertion of the polypeptide-encoding sequence will render the polyhedrin gene inactive and produce recombinant virus lacking coat protein. The recombinant viruses may then be used to infect, for example, S. frugiperda cells or Trichoplusia larvae in which the polypeptide of interest may be expressed (Engelhard, E. K. et al. (1994) *Proc. Natl. Acad. Sci. 91* :3224-3227).

In mammalian host cells, a number of viral-based expression systems are generally available. For example, in cases where an adenovirus is used as an expression vector, sequences encoding a polypeptide of interest may be ligated into an adenovirus transcription/translation complex consisting of the late promoter and tripartite leader sequence. Insertion in a non-essential E1 or E3 region of the viral genome may be used to obtain a viable virus which is capable of expressing the polypeptide in infected host cells (Logan, J. and Shenk, T. (1984) *Proc. Natl. Acad. Sci. 81*:3655-3659). In addition, transcription enhancers, such as the Rous sarcoma virus (RSV) enhancer, may be used to increase expression in mammalian host cells.

Specific initiation signals may also be used to achieve more efficient translation of sequences encoding a polypeptide of interest. Such signals include the ATG initiation codon and adjacent sequences. In cases where sequences encoding the polypeptide, its initiation codon, and upstream sequences are inserted into the appropriate expression vector, no additional transcriptional or translational control signals may be needed. However, in cases where only coding sequence, or a portion thereof, is inserted, exogenous translational control signals including the ATG initiation codon should be provided. Furthermore, the initiation codon should be in the correct reading frame to ensure translation of the entire insert. Exogenous translational elements and initiation codons may be of various origins, both natural and synthetic. The efficiency of expression may be enhanced by the inclusion of enhancers which are appropriate for the particular cell system which is used, such as those described in the Litreature (Scharf, D. et al. (1994) *Results Probl. Cell Differ. 20*:125-162).

For long-term, high-yield production of recombinant proteins, stable expression is generally preferred. For example, cell lines which stably express a

polynucleotide of interest may be transformed using expression vectors which may contain viral origins of replication and/or endogenous expression elements and a selectable marker gene on the same or on a separate vector. Following the introduction of the vector, cells may be allowed to grow for 1-2 days in an enriched media before they are switched to selective media. The purpose of the selectable marker is to confer resistance to selection, and its presence allows growth and recovery of cells which successfully express the introduced sequences. Resistant clones of stably transformed cells may be proliferated using tissue culture techniques appropriate to the cell type.

Any number of selection systems may be used to recover transformed cell lines. These include, but are not limited to, the herpes simplex virus thymidine kinase (Wigler, M. et al. (1977) *Cell 11*:223-32) and adenine phosphoribosyltransferase (Lowy, I. et al. (1990) *Cell 22*:817-23) genes which can be employed in tk.sup.- or aprt.sup.- cells, respectively. Also, antimetabolite, antibiotic or herbicide resistance can be used as the basis for selection; for example, dhfr which confers resistance to methotrexate (Wigler, M. et al. (1980) *Proc. Natl. Acad. Sci. 77*:3567-70); npt, which confers resistance to the aminoglycosides, neomycin and G-418 (Colbere-Garapin, F. et al (1981) *J. Mol. Biol. 150*:1-14); and als or pat, which confer resistance to chlorsulfuron and phosphinotricin acetyltransferase, respectively (Murry, *supra*). Additional selectable genes had been described, for example, trpB, which allows cells to utilize indole in place of tryptophan, or hisD, which allows cells to utilize histinol in place of histidine (Hartman, S. C. and R. C. Mulligan (1988) *Proc. Natl. Acad. Sci. 85*:8047-51). The use of visible markers has gained popularity with such markers as anthocyanins, beta-glucuronidase and its substrate GUS, and luciferase and its substrate luciferin, being widely used not only to identify transformants, but also to quantify the amount of transient or stable protein expression attributable to a specific vector system (Rhodes, C. A. et al. (1995) *Methods Mol. Biol. 55*:121-131).

Although the presence/absence of marker gene expression suggests that the gene of interest is also present, its presence and expression may need to be confirmed. For example, if the sequence encoding a polypeptide is inserted within a marker gene sequence, recombinant cells containing sequences can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with a polypeptide-encoding sequence under the control of a single promoter. Expression of

the marker gene in response to induction or selection usually indicates expression of the tandem gene as well.

Alternatively, host cells that contain and express a desired polynucleotide sequence may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridizations and protein bioassay or immunoassay techniques which include, for example, membrane, solution, or chip based technologies for the detection and/or quantification of nucleic acid or protein.

A variety of protocols for detecting and measuring the expression of polynucleotide-encoded products, using either polyclonal or monoclonal antibodies specific for the product are known in the art. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), and fluorescence activated cell sorting (FACS). A two-site, monoclonal-based immunoassay utilizing monoclonal antibodies reactive to two non-interfering epitopes on a given polypeptide may be preferred for some applications, but a competitive binding assay may also be employed. These and other assays are described, among other places, in Hampton, R. et al. (1990; Serological Methods, a Laboratory Manual, APS Press, St Paul. Minn.) and Maddox, D. E. et al. (1983; *J. Exp. Med. 158*:1211-1216).

A wide variety of labels and conjugation techniques are known by those skilled in the art and may be used in various nucleic acid and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting sequences related to polynucleotides include oligolabeling, nick translation, end-labeling or PCR amplification using a labeled nucleotide. Alternatively, the sequences, or any portions thereof may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes in vitro by addition of an appropriate RNA polymerase such as T7, T3, or SP6 and labeled nucleotides. These procedures may be conducted using a variety of commercially available kits. Suitable reporter molecules or labels, which may be used include radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents as well as substrates, cofactors, inhibitors, magnetic particles, and the like.

Host cells transformed, transfected, or infected, for example, with a polynucleotide sequence of interest may be cultured under conditions suitable for the expression and recovery of the protein from cell culture. The protein produced by a

recombinant cell may be secreted or contained intracellularly depending on the sequence and/or the vector used. As will be understood by those of skill in the art, expression vectors containing polynucleotides of the invention may be designed to contain signal sequences which direct secretion of the encoded polypeptide through a prokaryotic or eukaryotic cell membrane. Other recombinant constructions may be used to join sequences encoding a polypeptide of interest to nucleotide sequence encoding a polypeptide domain which will facilitate purification of soluble proteins. Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals, protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp., Seattle, Wash.). The inclusion of cleavable linker sequences such as those specific for Factor XA or enterokinase (Invitrogen. San Diego, Calif.) between the purification domain and the encoded polypeptide may be used to facilitate purification. One such expression vector provides for expression of a fusion protein containing a polypeptide of interest and a nucleic acid encoding 6 histidine residues preceding a thioredoxin or an enterokinase cleavage site. The histidine residues facilitate purification on IMIAC (immobilized metal ion affinity chromatography) as described in Porath, J. et al. (1992, *Prot. Exp. Purif.* 3:263-281) while the enterokinase cleavage site provides a means for purifying the desired polypeptide from the fusion protein. A discussion of vectors which contain fusion proteins is provided in Kroll, D. J. et al. (1993; *DNA Cell Biol. 12*:441-453).

Each individual regulatory sequence or active chromosomal element may affect regulation of protein and/or RNA expression. For example, methods of the invention were used to detect the DNaseI hypersensitive sites associated with the β-globin locus LCR, which is known to have a regulatory function, as described in Li *et al.*, 1999 *Trends Genet.* 15:403. The controlled gene (including the coding sequence and any regulatory sequences adjacent to the coding regions) in many or most cases will be adjacent (within one base pair away) or close to (within 10, 100, 500, 2000, or even 10,000, 20,000, 50,000 or 100,000 base pairs) the regulatory sequence, although in some circumstances it may be farther than 100,000 base pairs from the regulatory sequence.

Libraries comprising RSs, regulatory sequences and/or regulatory units of the invention are included in the present invention. Libraries of polynucleotides having sequences identified as RSs, particularly within a certain profiled loci, are useful for a variety of purposes, including, for example, for identifying sequences that coordinately regulate specific genes or sets of genes. Sets and subsets of the sequences listed in the figures have particular value in embodiments of the invention. A library comprises at least two polynucleotides of the invention or fragments or functional fragments thereof or cells comprising the same. Libraries may comprise isolated nucleic acid fragments, vectors comprising inserts corresponding to nucleic acid sequences of the invention, or cells comprising such vectors, for example. Other embodiments of libraries and sets utilize both sequences and locations together. A library or "set" as termed here may have at least two members, but more preferably has at least 10 members, 100 members, 500 members, 1,000 members, 2,000 members, 5,000 members, 10,000 members, 20,000 members or even more than 30,000 members. A particularly desirable embodiment provides a set of members (sequences, position location or both) of regulatory units associated with chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,18, 19, 20, 21, 22, or 23. Another embodiment of the invention provides a set of regulatory units associated with gene expression in a particular cell or tissue (*e.g.* a differentiated cell or a diseased cell), at a specific developmental stage, or in response to an external stimuli (*e.g.* hormone, cytokine, chemical, small molecule, UV, or radiation).

The present invention also provides arrays and microarrays of polynucleotides of the invention or fragments or functional fragments thereof. In addition, arrays may comprise cells of the invention. In certain embodiments, such arrays comprise two or more different polynucleotides or cells, each located at a discrete and identifiable position on a solid support or in discrete vessels. In another embodiment, an array may comprise a plurality of different polynucleotides or cells, with several different polynucleotides or cells located at a discrete and identifiable position on a solid support or in a discrete vessel. In one embodiment, an array comprises a plurality of regulatory units. In a related embodiment, an array comprises a plurality of regulatory sequences grouped according to the regulatory unit with which they are associated. Thus, an array may comprise, for example, discrete vessels or positions, each containing a regulatory unit or regulatory sequences associated with a regulatory

unit. Preferably, each position or vessel comprises between 3 and 10, or between 10 and 100 different polynucleotides or cells.

5.10 Databases and Computer Systems

Information regarding the identity and location of RSs, particularly in the context of regulatory sequence profiles, has great utility and may be used in several ways. At least some or all of the information may be assembled as a database within a storage unit, such as a computer readable medium. The information may be transferred in electronic form, such as over the Internet or over a telephone line. The information has great value for diagnostic or identification purposes or for studying regulatory systems, and may for example be used as a resource for selecting a desired regulatory sequence. Accordingly, the invention provides databases that comprise RSs and/or RS profiles. Such databases may, for example, be used for data mining. The locus information shown in these figures show not only useful active chromatin element, which likely correspond to regulatory sequences, but also reveal details of how to position the sequences with respect to genes encoded by open reading frames, which are regulated by the regulatory sequences. Accordingly, in an embodiment a data set provides the sequence of a new regulatory sequence along with positional information needed to utilize the regulatory sequence in combination with other genes.

A database may comprise as little as a single locus that shows the sequence of a regulatory sequence and positional information. A database may include multiple members of regulatory sequences and preferably contains enough members to allow use of the database as a library reference, whereby a selection may be made among two or more regulatory sequences. Thus, a database, as termed here often contains at least two sequences. In many embodiments, a database will include many loci information. In certain embodiments, a database includes and identifies one or more members of a regulatory unit associated with a loci or gene. In certain embodiments, a database includes relative positional information for one or more regulatory sequences of a regulatory unit. In a further embodiment, a database may include relative positional information for one or more regulatory sequences of a regulatory unit and an associated or regulated gene. In another embodiment, a database includes regulatory sequences associated with one hypersensitive site or region of genomic DNA.

One or more databases or computer readable media may be operated on by a stored program. Stored programs including information regarding regulatory sequences and units may be used in any of a variety of manners, for example, to gain information regarding sequences to use to control expression of a gene. A stored program may be used to couple a desired regulatory sequence with a gene sequence whose regulation is desired. The stored program may select the coupled regulatory sequence based on a variety of different factors, including, for example, its genomic position relative to the gene to be regulated, correlation of a desired expression pattern with a particular gene and associated regulatory sequence and/or regulatory unit, and positional information related to the regulatory sequence and an endogenously regulated gene. Accordingly, an embodiment of the invention provides a program that stores both sequences in memory and combines the sequences in a meaningful way that may be used to design and carry out genetic manipulations.

Other manipulations are possible by a stored program, such as direct sequence comparisons between regulatory sequences and comparisons between stored regulatory sequences and other sequences that may be inputted for a desired comparison. Accordingly, a stored program may be used to identify regulatory sequences and consensus sequences and to identify regulatory sequences within a gene, as well as functional relationships between different regulatory sequences. For example, the presence of two or more of the same regulatory sequences within different regulatory loci suggests that these sequences function cooperatively to regulate gene expression.

Stored programs may similarly be used for comparison of regulatory sequence profiles generated in different cells, for example.

Software programs are contemplated for discovery and use of regulatory sequences, RSs, and positional information. In each case, a set of regulatory sequences, regulatory units, and/or positions in the human genome are loaded into a computer and stored, in volatile memory, short-term erasable memory and/or long term non-erasable media. The set may comprise an regulatory sequence profile. A program is loaded into the computer that parses through the set of sequences and/or genomic positions. For each parse, the computer makes a decision having biological or biochemical relevance. For example, one type of decision is to determine whether a parsed sequence is similar to (homogenous to) a known active genetic sequence such as a known promoter or so-called "enhancer" sequence. The computer may look

for strict equivalency in sequence of course but in many embodiments the computer will examine for a minimum percent homology or other correspondence as is known in this art. By way of example, if a segment of a sequence of about 15, 20, 30, 50, 75, 100, or 200 bases of those sets of genome locations of regulatory sequences is at least 70%, 75%, 80%, 85%, 90%, 95%, or at least 97% identical to a reference known active genetic sequence, the computer will store the correspondence information or match in memory for use later by a program or for display to the computer operator. In most embodiments, the computer will store selections in memory and later transmit a set of selections by electronic transfer to a permanent medium such as an optical or magnetic disk or by electronic transmission.

Sets of sequences and/or positional locations may be prepared by computer analysis of information from sets of genome locations of regulatory sequences, and have great intrinsic value for a variety of uses such as regulatory unit discovery, diagnostics and therapeutics. In many embodiments, a computer program is used that inputs at least part of a regulatory sequence or RS, such as at least 10, 100, 1,000, 10,000 or more sequences and genome locations and then selects out a smaller set therefrom. Particularly contemplated are sets of sequences and/or genome positions that correspond to regions of the genome, such as particular chromosomes, hypervariable regions that experience high levels of DNA breakage, and the like. After computer formation, such sets of data, presented in computer readable form or directly readable by a person, are valuable items of commerce and may be sold directly.

The DNA sequences and their location information shown in the figures may be used for further discoveries through data mining, using a portion or all of the listed information. For example, the figures reveal coordinate expression of regulatory active genetic sites in genome space, as can be readily apprehended by a computer directed by a program to group regulatory sequences from the figures that physically locate close to each other in the genome. Generally speaking, such grouped sites, termed "clusters" herein for convenience, regulate coordinately one or more genes. Such clusters are regulatory units.

In a desirable embodiment, a software program instructs a computer to load multiple genome locations of regulatory sequences and then compare how far apart each genome location is from the others. The program instructs the computer a set maximum genome distance for comparison and to decide if two sites are less apart

than that distance. If they are, then this fact is noted in memory, such that the two are labeled or grouped into the same cluster. Most conveniently, a cluster will be made by storing identifiers for the regulatory sequences at the same or adjacent areas of memory. Cluster groups may be stored on long-term media (*e.g.*, hard drive, CD ROM) and/or displayed to the computer operator. In an embodiment, two regulatory sequences are deemed within the same cluster if their genome locations are within 1,000,000 bases of each other. In another embodiment, regulatory genetic sites are deemed part of a cluster if their genome locations are within 300,000 bases, 100,000 bases, 30,000 bases, 10,000 bases, 3,000 bases 1,000 bases or even 250 bases of each other.

For purposes of brevity, a separate listing of each possible subset of sequences contemplated is not presented, and space limitations are overcome by the convenient use of computers to group the data as described herein. Thus, specific clusters of regulatory sequences, found on each chromosome and selected by closeness based on proximity are intended embodiments and can be easily printed in tabular form as desired with the aid of a computer.

One embodiment of the invention provides a computer program that determines a cluster by reviewing the genome positions of multiple regulatory sites (at least 100, 1,000, 10,000 or more) and placing sites having near positions to each other within one or more of the above specified ranges, into a common group. This embodiment of the invention is made possible by the fact that the information displayed in the figures were obtained under real conditions wherein multiple coordinating active genetic sites were actively controlling gene expression. That is, the sites listed in this figure are not a random assortment of active genetic sites in the genome and do not necessarily represent all possible sites, but represent active genetic sites that were simultaneously active in a functioning cell system. Among other things, this property distinguishes the information of the figures from other data sets obtained by others using purely computer analysis of the sequenced genome.

The invention further includes the information gathered by the previously described computer programs, including, for example, the genomic position of identified sequences. Such information includes data and data sets in both printed and computer-readable format. Thus, the invention provides computer readable medium comprising a plurality of polynucleotide sequences identified as regulatory sequences or regulatory units. The data may be further defined or classified to include

regulatory sequences or regulatory units associated with gene expression or alterations in gene expression in specific cells or tissues, diseases, chromosomes, transcription factors, or chromatin structure or modifications, for example. Indeed, the skilled artisan would readily appreciate that data or computer-readable medium containing hypersensitivity sites or active control elements from any cell of interest, including, for example, a cell treated with a drug or drug candidate, is contemplated by the invention and of value in identifying genes and gene regulation associated with the cell.

The loci information may be also used for a business method, such as, for example, a method of adding biochemical regulation to a known gene having a commercially valuable function. Such method may, for example, involve selecting a known gene having a desirable commercial value, selecting a regulatory sequence from a database that contains one, two or more regulatory sequences; operably linking the selected regulatory sequence with the selected known gene to form a polynucleotide comprising both; and commercially using the nucleic acid by, for example selling it, transforming an organism to express the protein and thereby increasing the commercial value of the organism, or making a vector having a commercial value.

The specific sequences and precise genomic positions of RSs with respect to cognate genes may be modified and used in combination with other nucleic acid, as will be readily appreciated by skilled artisans. For example, the sets of genome locations of regulatory sequences may be used in whole or in part by itself or linked in combination with other sequences as discovery tools and medical agents.

In the context of arrays of nucleic acid sequences of the invention, a computer-readable medium may include any of the information described *supra*, and, in addition, may further comprise the array location of each arrayed nucleic acid sequence. The computer-readable medium may, therefore, provide the sequence of a nucleic acid located at a specific location on an array. The computer-readable medium may further provide any other known information about the sequence at a specific location on an array, including, for example, the genomic location of the sequence, any genes associated with the sequence, the regulatory unit associated with the sequence, other sequences within this regulatory unit, transcription factors associated with the sequence, and any diseases or disorders associated with the sequence, *etc.*

As another example, a software program can direct a computer to find allelic forms of a regulatory sequence or regulatory unit by searching public databases for sequences of regulatory sequences from information of genomic location that may be input into the computer. In certain embodiments, the allelic variant may be identified based upon its having the same genomic position a an identified regulatory sequence or regulatory unit. In another embodiment, a computer under direction of a program inspects the genome location contents of a database and chooses a regulatory genetic site near a desired gene, thereby determining a previously unknown regulatory sequence or unit, a new function for an under appreciated functionally regulatory sequence or unit, or may provide greater clarity as to the borders of a known regulatory unit. According to this embodiment, preferably the computer looks for regulatory sequences within 100,000 base pairs of a selected gene, and more preferably within 50,000; 20,000; 10,000; 3,000; 2,000; 1,000; 500; 200; 100; or even 50 base pairs of a given selected gene.

The DNA sequences and their location information shown in the figures may be used for further discoveries through data mining, using a portion, or all of the listed information. For example, the figures reveal regulatory sequences that coordinately regulate gene expression as a regulatory unit.

Another desirable embodiment provides one or more sets of members (either polynucleotide, sequences, genome locations or both) that are associated with genetic abnormalities of uncontrolled cell growth. A skilled artisan may prepare a set of genetic anomalies associated with one or more human diseases (*e.g.*, cancers, immune disorders, neurological disorders, cardiac disorders). For example, by matching known genetic changes associated with malignant transformation with the precise sequence or position information of a regulatory sequence or regulatory unit, it is possible to identify genetic anomalies associated with a specific cancer. In one embodiment, a pre-existing set of genetic changes associated with a disease are compared to determine which of the changes linked to disease involve regulatory DNA. This information provides great value for drug discovery and for new modalities for treating disease.

In a related embodiment, a software program instructs a computer to compare known sets of genetic changes associated with a disease with RSs, regulatory genetic sequences and/or regulatory units. The computer inputs at least one set of genetic information, inputs at least some sequence information and or genome positions and

compares identities using a known algorithm or procedure. After comparing the two sets, the computer selects a match set to be output or used for further analysis, indicating one or more sequences associated more definitively as active genetic regions of more defined sequence and size.

Regulatory sequences and regulatory units may be prepared and used as articles of commerce, research tools, diagnostic aids, drug discovery aids and the like, based on a desirable grouping category such as those based on genetic changes in malignancy and genetic changes associated with specific disease. In a very desirable embodiment, a known genetic abnormality is used to find linked regulatory sequences and/or regulatory units that cooperatively influence gene expression or an overall biological process mediated by one or multiple genes. This is carried out by examining for unknown cluster partners. In this embodiment, RSs, RS profiles or regulatory sequences that associate with a known DNA problem, such as a disease or allelic form of a gene associated with a definable trait based on, for example, an improper transposition, deletion, or other mutation, are placed into a set and combined with further members that are found to be associated with the known genetic errors.

The invention further includes software that performs any or any step or aspect of the methods of the present invention described *infra*.

5.11 Methods Using Regulatory Sequences and Units

The sequences and information of the invention may be used in a variety of methods, including, but not limited to, methods related to the identification of genes, the regulation of gene expression, and diagnostic and therapeutic methods involving regulatory sequences and regulation of gene expression.

The information presented in or obtained from the figures herein may be used to identify or derive new, previously unknown regulatory sequences and regulatory units for known genes. In one embodiment, a regulatory sequence for a gene is discovered or further characterized by comparing the positional information of the seqeunce with the known location of the gene. The position of a known gene may be compared to positional information of regulatory sequences and units to identify a regulatory sequence or unit near the gene (preferably within 50,000, 20,000 or10,000 base pairs, more preferably within 3,000, 1,500 or 500 base pairs or any integer value between). In certain embodiments, the ability of the identified regulatory sequence to regulate the known gene may be further confirmed by functional experiments, such as

site-directed mutagenesis of the identified regulatory sequence, *e.g.* knocking out the sequence in a cell or animal.

Furthermore, sequence information obtained from the figures may be used to design primers for polymerase chain reactions (PCR). A regulatory sequence or unit that is close (preferably within 10,000 base pair, more preferably within 3,000, 1,500 or 500 base pairs) to a gene, single nucleotide polymorphism (SNP), or other site of interest, may be selected by a computer. Sequences for primer recognition can be selected and PCR reactions performed to identify and/or quantitate SNPs, changes in chromatin structure, or genome-specific mutations or individual-specific mutations.

In one embodiment, a gene already has a known regulatory sequence that may be similar in location to a second regulatory sequence. This information may be used to discover a further attribute of the known regulatory sequence or unit, such as the location of a regulatory sequence that may be at the edge or outside the known regulatory unit. In the latter case, this embodiment of the invention allows the discovery of a new section or border of a previously considerer regulatory unit. Thus, the invention includes methods of identifying regulatory units comprising two or more regulatory sequences.

In a related embodiment, multiple regulatory sequences or regulatory units that affect the same gene or set of genes are discovered by virtue of their clustering in genome space. Such clustering may be based on their physical proximity to each other, their location on the same chromosome, or their physical proximity to the same gene. Clustered regulatory sequences and units are preferably within 10,000 base pair, more preferably within 3,000, 1,500 or 500 base pairs of each other.

Regulatory sequences and units associated with a specific phenotype, such as, for example, a disease or disorder, a differentiated cell type, or a specific developmental stage, may be identified according to the invention, for example, by identifying a gene expressed in a cell displaying the phenotype, identifying the genomic position of the gene, and identifying a regulatory sequence or unit located near the gene. Such regulatory sequences may then be used to direct expression of a gene in a similar manner as the original gene associated with a particular cell type or phenotype. This is particularly useful for targeting gene expression to a specific cell type or at a specific developmental stage, for example, in gene therapy methods designed to deliver a therapeutic nucleic acid or polypeptide, described *infra*.

The invention further allows the identification of a gene associated with a disease or disorder resulting, for example, from faulty regulation of gene expression. In one embodiment, a regulatory sequence or unit of the invention is identified as being mutated in a cell or patient suffering from a disease or other phenotypic or biochemical disorder by any available means, including sequencing or examination of either chromatin structure or associated polypeptides, for example. The location of the mutated regulatory sequence or unit is identified and a gene located nearby is further identified. In certain embodiments, expression of the gene in normal cells versus cells with a disease or disorder is compared to confirm that the gene is expressed differently in cells with a disease or other disorder. Methods of comparing mRNA levels are widely known and routine in the art, including, for example, northern blotting, RT-PCR and real-time PCR.

The databases and profiles of the invention may be used for discovery and analysis of DNA-protein interactions. The identities of proteins participating in the interactions and their functions may be determined. For example, key proteins involved in transcriptional regulation may be identified. In one embodiment, regulatory sequences or units, or RSs are labeled and used as substrates in electro-mobility shift assays (EMSA) to identify which proteins from a range of nuclear extracts bind to the sequence. Addition of antibodies raised against candidate nuclear proteins can be used to cause a further supershift allowing identification of the individual protein components within the nucleoprotein complex. Components may also be identified by direct sequencing of identified polypeptides or cloned polynucleotides encoding the polypeptides. Once the nature of the proteins are known the presence of post-transcriptional modifications can be determined, by use of antibodies raised against specific modifications or by mapping the mass of fragments by mass spectrophotometry, for example. It is understood in the art that the formation of transcriptional regulatory complexes may require cooperative interaction between polypeptides binding at two or more regulatory sequences. Accordingly, the use of regulatory units comprising more than one regulatory sequence is advantageous in that it allows the isolation and identification of additional components of an associated regulatory complex, which may include polypeptides that do not directly bind to a regulatory sequence, including those whose recruitment required cooperative binding of two or more polypeptides to discrete regulatory sequences.

The databases may be used as templates for *in vitro* or *in vivo* footprinting and identify the positions of DNA-binding proteins. 'Footprinting' of the cloned sequences may be carried out with a variety of cutting agents, such as DNaseI or free radicals for example. This reveals patterns of binding of proteins either *in vitro* to a panel of nuclear extracts or purified components or *in vivo* in different tissues. The binding of a particular protein is specific to its cognate site, many of which are known and hence can be used to infer the proteins bound to the regulatory site. The region of the regulatory site that the protein covers can indicate the overall structure, and therefore function, of the regulatory site.

The databases can identify proteins bound to and associated with regulatory sites. The identification of all of the components of the regulatory sites *in vivo* is made possible by hybridizing nucleic acid having sequence(s) of cloned regulatory sites to exposed regions of fractionated chromatin. For example, cross-linked sonicated chromatin can be treated with exonucleases to expose single-stranded DNA regions that can form targets for biotinylated fragments from the cloned regulatory sites. Such captured complexes can be analyzed for protein content and levels of epigenetic modifications. In this example both protein-DNA and protein-protein interactions can be determined. The available techniques for carrying out these studies potentiate the discovery of interactions between regulatory sites, as proposed by looping models wherein transcriptional enhancers interact with their cognate promoters via complex protein-protein interactions. The existence of such complexes may be a general effect or may be restricted to a number of super-regulatory elements or LCRs (Locus Control Regions).

Another desirable embodiment of the invention is to utilize isolated fragments, containing regulatory sequences and/or units and preferably labeled with a detectable label, (i) to probe to find and identify complementary genomic DNA sequences and (ii) to find and identify proteins and protein complexes with binding affinity for regulatory sites. Suitable techniques include cross-hybridization; immunoprecipitation and related antibody-based methods; cross-linking and related methods – all well known to those of ordinary skill in the art. These embodiments of the invention can, for example, detect new DNA binding proteins, reveal families of regulatory regions and mechanisms, and provide new modalities for controlling transgenic expression.

A library of isolated fragments can be used for genome-wide location profiling of DNA-bound proteins. In one embodiment, dynamic binding of gene-specific transcription factors and components of the general transcription apparatus is monitored in cells. In one such embodiment, yeast cells are used to determine very broad based fundamental proteins shared between yeast and mammals. In another embodiment, mammalian cells are used to determine regulatory proteins involved in mammals. In yet another embodiment, both yeast and mammalian cell results are compared to group the regulatory proteins accordingly.

The genome-wide location method can correctly identify known sites of action for transcriptional activators and reveal unexpected functions for these activators. In one embodiment, the tools taught in U.S. 6,410,243 issued June 25, 2002 are used. In another embodiment, a library of isolated fragments is obtained from cells in each of at least two conditions. The combination of condition expression and DNA location profiles obtained by library comparison can identify a global set of genes whose expression is under the direct control of specific activators and components of the transcription apparatus as cells respond to changes in their extracellular environment.

In another embodiment, sequences are used for raising antibodies against components of isolated regulatory complexes. Successful isolation of the intact nucleoprotein complexes by hybridization with sequences derived from the cloned regulatory sites allows the generation of monoclonal and polyclonal antibodies against both the proteins bound in the complex and the tertiary structure of it. Such antibodies are useful in a range of applications such as CHIP, wherein antibodies raised against the nucleoprotein complex as a whole have higher specificity. The antibodies also may be used in studies that disrupt the function of the regulatory site by interfering with molecule(s) that interact with the regulatory site in its natural context.

The databases further may be used as substrates for producing cross-referenced libraries to define key active genetic elements. Many regulatory sites are common between tissues and cells exposed to different stimuli. For example, some regulatory sites are associated with constitutively expressed genes, and some are unique and define the cell and its transcriptional program. To find these differentially formed regulatory sites, subtracted libraries can be made using regulatory sites cloned from two different populations as substrates. Methods of producing subtracted libraries are known in the art.

The databases further may be used to identify regulatory elements and units in various organisms. Databases of regulatory sites can be used to search for homologues from different organisms and in this way identify regulatory sequences, units, and relationships in other organisms.

Yet another use of the databases is for the study of post-transcriptional modifications within the genome. The CHIP protocol (chromatin immuno-precipitation) has been used to enrich for sequences, often from formaldehyde cross-linked nuclei, bound by nuclear proteins or by proteins carrying post-transcriptional modifications, such as the acetylation pattern of histones. This pool of fragments can be used to hybridize to the isolated regulatory sequences and units to determine, for example, which regulatory sites are bound by which nuclear protein. In the case of post-transcriptional modification, the changes in these epigenetic markers can be followed as a function of tissue-type and development, for example.

The databases may be used to probe the role of differential methylation within active genetic elements. Analysis of the sequences of the cloned regulatory sites can reveal the presence of CpG-dinucleotides. Some of these dinucleotides can be differentially methylated at cytosines,and such methylation may causes transcriptional inactivity of an associated gene. Genomic sequencing can be used to compare the methylation status of such potential epigenetic modifiers across a panel of nuclei to identify those that have key regulatory functions.

The databases may also be used for stimulating methylation at regulatory sites by introducing a complementary and methylated polynucleotide. At active genetic elements marked by regulatory sites, a strong correlation exists between demethylation of certain sites (of the cytosine in CpG dinucleotides) and transcriptional activity. These key CpG dinucleotides can be re-methylated by introduction of a complementary polynucleotide containing a 5-methylcytosine at the crucial position; the resultant hemimethylated site will be a substrate for the maintenance methylase activity present in eukaryotic cells. The introduction of a methylated CpG dinucleotide into the active regulatory sequence would be expected to change its transcriptional influence.

The sequences may also be used as markers for studying the role of nuclear localization in transcriptional induction. It is possible to follow the nuclear localization of specific sequences using fluorescently labeled probes and confocal

microscopy. The existence of sub-compartments within the nucleus and the recruitment of active genetic sequences and genes to them potentially plays a major role in understanding transcriptional regulation in eukaryotic nuclei. Most preferably, a panel of labeled probes is generated against a regulatory unit or sets of regulatory sequences. The distribution of the unit or sequences may be monitored throughout the nuclei and, in certain embodiments, compared with different systems or under different conditions. The invention, therefore, provides a method of determining the nuclear localization of regulatory sequence and units of the invention comprising preparing one or more labeled probes corresponding to an RS or regulatory sequence, introducing said probes to a cell, and determining the subcellular localization pattern of the probes at one or more different times.

The invention provides a variety of methods of regulating gene expression. In one embodiment, gene expression is regulated by preparing a polynucleotide comprising an RS, a regulatory sequence or regulatory unit, or a fragment thereof, and a gene. The polynucleotide may comprise one or more than one regulatory sequence or regulatory unit. The invention thus provides methods of targeting gene expression, for example, to a particular cell type or during a particular developmental stage, which comprise identifying a regulatory sequence or regulatory unit associated with a desired pattern of gene expression, preparing a polynucleotide comprising the identified regulatory sequence or regulatory unit and the gene to be expressed, and introducing said polynucleotide into a cell. The regulatory sequence or unit may be identified based upon its endogenous or genomic physical or functional association with a gene that is expressed in the desired manner. In certain embodiments, the method may be used to prepare a transgenic animal. In other embodiments, the method may be used, for example, in gene therapy, to provide a therapeutic molecule. The therapeutic molecule may be of any form, including RNA or polypeptide, for example, and may act in any therapeutic manner, including inhibiting or enhancing the expression or activity of a polypeptide or providing a polypeptide that is not normally expressed at normal levels in a cell.

RSs, including those identified in RS profiles, have a very wide range of uses. Embodiments of the invention may utilize one, two, multiple sequences or even all sequences each of these uses as may be desired. The sequences may be used in screening for the formation of regulatory sites by hybridization. That is, the sequences of the RSs can be used as a substrate for screening by hybridization, by

immobilization on a nylon filter, or glass microarray, for example. The generation of probe populations, enriched or depleted in fragments covering regulatory sites, from different nuclei preparations can allow determination of which regulatory sites are structurally present. Correlations can be drawn between RSs specific for different tissues, developmental stages, application of stimuli or disease state, for example.

The identification and characterization of the regulatory sequences, for example, by RS profiling, allows the identification and determination of regulatory sequences and units active or associated with specific patterns of gene expression, *e.g.* in differentiated cells, at certain developmental stages, or in response to stimuli. As indicated *supra*, the invention includes programs that compare the active regulatory sequences and units identified in different cells and databases comprising such information. In addition to these methods of identifying regulatory sequences and units associated with a specific pattern of gene expression, the invention further provides methods of sorting or classifying cells based upon their active regulatory sequences or units. Such methods typically compare the active regulatory sequences or units of a sample cell to those of one or more other cells and determine which other cell has the same or most similar pattern or profile of active regulatory sequences or units as the sample, thereby determining the cell type of the sample cell. Active regulatory sequences may be identified by any means available in the art, including, for example, hypersensitivity site mapping, chromatin structure analysis, and identification of associated polypeptides, some of which are described in detail in U.S. patent applications No. 60/108,206, No. 09/432,576, No. 60/302,369, No. 60/290,036, No. 60/294,890, No. 60/294,890, No. 60/378,664, No. 60/387,910, No. 60/387,887, No. 10/187,887, and No. 60/404,121, and PCT applications PCT/US02/15032 and PCT/US02/16967, which are hereby incorporated by reference in their entirety.

Polynucleotides, information, and databases of the invention may be used as *in vivo* markers for classification and sorting of cells. In one embodiment, a cell of unknown origin may be classified based upon the active regulatory sequences or sites. The method may be used, for example, to determine the cell type or origin of metastatic or circulating tumor cells. In a related embodiment, the invention provides a method of determining whether a cell has been exposed to a particular agent, such as a drug, for example, by comparing the regulatory site profile of the cell to the

regulatory site profile of a similar cell wither treated or untreated with an agent. Thus, the invention provides a genetically-based drug test.

In addition to determining the status of a cell, these methods allow the sorting and isolation of cells with specific identifiable active regulatory loci. For example, the formation of certain regulatory sites crucial for induction of certain genes may define the position at checkpoints of each cell in terms of its developmental progress and tissue specificity. Using labeled probes directed towards the accessible regions of regulatory sites, which remain inaccessible when the site is not formed, allows the detection of such 'markers' in intact nuclei. By using, for example, two fluorescently-labeled probes that give a strong FRET signal when bound to the same region of a formed regulatory site, it is possible to fractionate (using FACS) a population of cells from complex mixtures according to their exact developmental stage or tissue specificity.

The databases may similarly be used for functional tissue typing. The ability to detect formation of regulatory sites in nuclei allows construction of a regulatory profile for mixtures of tissue, either separated from primary tissue or from monocultures. A thorough understanding of how these profiles change due to a stimulus, such as drug treatment, allows the isolation of cells from a previously homogenous population that are highly potentiated. An example is the sorting of totipotent stem cells from a larger population or stem cells that have successfully been pushed down a particular developmental pathway.

Sequences and sites identified using one or more of the sequences listed in the figure can be used in the diagnosis and treatment of diseases and disorders. A diagnostic may comprise a sequence, *e.g.* comprising a regulatory unit (which may be DNA, RNA or PNA), coupled to a solid support for the detection of a complementary sequence (which may be DNA or RNA). Levels of expression (or trends of expression over time) of the complementary sequence can be determined from a biological sample obtained from a patient (*e.g.*, DNA, hnRNA, mRNA, rRNA, miRNA, ncRNA, stRNA, RNAi) as a disease indication. Expression levels significantly lower or higher in the test sample as compared to expression levels in a normal control sample may indicate the presence of a disease or disorder. In certain embodiments, a reference value is determined based on the expression levels in one or more normal controls, and the presence of a disease or disorder is determined by comparing expression levels in the test sample to the reference value. In certain

embodiments, a two-fold difference in expression is considered significant, while in other embodiments, a three-fold, four-fold, or five-fold difference is considered significant. The skilled artisan would readily appreciate that normal levels of mRNA and polypeptide expression may fluctuate or vary between different normal controls, and will take such variation into account when determining an appropriate reference value and a significant level of variation from a normal value. Treatment may comprise therapeutic compositions of polypeptides (with one or more antigenic determinants) or antibodies (*e.g.*, monoclonal, polyclonal, humanized, or fragments of antibodies such as, for example, Fv fragments) to such polypeptides identified using the sequences of the invention.

The databases may be used in a variety of diagnostic methods. Correlations between a disease or other disorder and the activity of one or more regulatory sites or units of the invention may be established, and a disease or disorder correlated with a particular activity may be detected by examining the activity of the identified site in a subject, particularly a subject suspected of having the disease or disorder. Heretofore, a suitable comprehensive approach to discovery and intelligent manipulation of regulatory mutations on a regulatory unit genome wide basis there has not been available. In certain embodiment, libraries of the invention, both standardized and those obtained from specific diseased individuals may be obtained, and compared to detect regulatory changes associated with mutations.

A variety of mutation discovery techniques may be employed in these embodiments of the invention. One of the earliest methods detects restriction fragment length polymorphisms (RFLPs) using the Southern blotting technique. (Southern, E. M., *J. Mol. Biol.* 98:503-517 (1975)). RFLPs determine genetic variations in certain DNA fragments by cleaving the fragments with a type II restriction endonuclease. The differences in DNA length are due to the presence or absence of a specific endonuclease recognition site(s) and are detected using DNA hybridization with DNA probes after separation by gel electrophoresis. Other methods use polymerase chain reaction (PCR) techniques to detect sequence differences. In instances where a particular mutation has been identified, labeled primers can be used to determine whether a sample contains the known mutations. PCT/US93/04160 describes a method that allows perfectly matched DNA molecules to be separated from imperfectly matched molecules. The molecules can also be labeled to provide probes for identifying regions of heterozygosity in the genome.

Newer methods such as that described in U.S. 6,297,010 issued October 2, 2001 also are useful for finding specific changes in large-scale systems. Each of these systems specifically is contemplated for embodiments of the invention that utilize libraries of DNA fragments as taught therein.

In a particularly useful embodiment made possible by the invention, multiple identical regulatory sequences or RSs found at disparate regions of the genome are compared with each other and a mutation in one is determined. Unlike most other methods which either recognize only a single regulatory sequence or group of sequences, this embodiment of the invention allows the detection and response to the existence of allelic forms of the same regulatory unit as found in different locations in the genome. This embodiment may conveniently be implemented in large-scale systems using tools known in the art. For example, WO 95/12689, assigned to GeneCheck, Inc., describes contacting labeled heteroduplexed DNA with a labeled immobilized mismatch binding protein ("MBP") such as MutS. Binding, detected by direct or indirect methods, is indicative of a mismatch. This method, for example, may be implemented in an array of sequences to indicate the presence or absence of a mismatch. Similarly, WO 93/02216, assigned to Upstate Biotechnology, Inc. describes how mutations can be detected using a labeled antibodies specific for mismatch binding proteins to determine if a mismatch is present.

In another embodiment, a reference library of known regulatory sequences, such as an immobilized array of known sequences is used to find a regulatory mutation using mismatch binding proteins. A typical use of mismatch binding proteins is seen, for example, in WO 95/29258. In this case, library test strands of nucleic acids comprising regulatory sequences or units are hybridized to sample strands. The formed duplexes are contacted with a mismatch binding protein and the complex is then treated with an exonuclease. The digestion of the nucleic acid terminates at the position of any bound MBP. The relative sizes of the resulting degradation products are analyzed, for example, by electrophoresis, to determine the presence and approximate location of the mismatch.

Many regulatory sequences were discovered in close positional association with each other. That is, two or more sites are sometimes found and expressed in a common locus or regulatory unit that generally encompasses up to 1000, 10,000, 100,000, or up to 1,000,000 continuous base pairs of the genome. Such regulatory loci may comprise one or more genes. Without wishing to be bound by any one theory of this

embodiment of the invention, it is thought that coordinately expressed regulatory sites are placed in a common locus, for improved regulatory efficiency and coordinate activity. It was discovered that the diagnosis and intervention of a loci-organized regulatory system, in many cases requires simultaneous review and treatment of the member regulatory sequences for the loci system. Accordingly, embodiments of the invention generate, manipulate, analyze and use regulatory loci of genomic regulatory sequences that both display nuclease hypersensitivity sites and positional association.

In one embodiment, a disease of regulation is detected by examining a particular locus to determine whether all known members *(i.e.* genes) of a regulatory locus group of regulatory sites are expressed. If a member is not expressed, or if a member is expressed more readily than the others by comparison to that of a reference, then that regulatory locus site is deemed altered or deficient. By associating specific genetic regulatory elements of a locus with a known regulatory effect, more powerful diagnoses and treatments are available.

A surprising discovery in the context of regulatory locus systems is how many were found during practice of the invention, particularly in view of severely limited number of loci systems previously known. In an embodiment of the invention, clusters of regulatory sequences are detected and manipulated as super-regulatory elements known as locus control regions (LCRs), which are capable of both regulating chromatin structure over large distances and enhancing transcription of a family of genes. A wide variety of techniques generated by others for studying LCRs are specifically contemplated for use in these embodiments of the invention. Methods had been developed for evaluating the twenty-five LCRs that had been identified in humans (Li *et al.*, 2002. *Blood* 100, 3077-3086), including the β-globin locus. The human version of this LCR consists of four tissue-specific DNaseI-regulatory sites, HSI to HS4, within a 25 kb region upstream of the five β-globin-like genes (reviewed in Fraser and Grosveld, 1998. *Curr. Opin. Cell Biol.*. 10, 361-365).

Many of the sequences presented herein are associated with regulatory sequences involved in disease. Labeled (by fluorescent dyes for example) polynucleotide probes (such as complementary PNA) or synthetic molecules designed to recognize stretches of the highly accessible regions of the DNaseI or hypersensitive regulatory sequences can be used to detect their formation in intact nuclei. The detection of those regulatory sequences associated specifically with disease states

either by studying nuclei which had been isolated or are still intact could be used to detect, evaluate and monitor cells with a regulatory environment associated with a disease. Changes in regulatory sequences associated with a disease or disorder may also be detected by other means, including, for example, sequencing.

In certain embodiments, nucleic acids of the invention may be used to identify a gene associated with a disease or disorder. For example, hypersensitivity sites present in a diseased cell or tissue may be identified and compared to those present in a normal cell or tissue. Hypersensitivity sites either present or lacking, or present to a differing degree, in the diseased cell or tissue are considered associated with the disease or disorder. The genomic location of one or more differing hypersensitivity sites may be determined and a gene located near to the site or known to be regulated by the site may be identified as involved in or associated with the disease. One of several means of confirming the relationship of the identified gene and the disease is to measure the levels of mRNA or polypeptide expressed from the gene and compare it to the levels observed in normal cells. Any different would confirm that the gene is associated with the disease or disorder. Methods of measuring mRNA and polypeptide levels are widely known and available in the art and include, for example, RT-PCR and western blotting, for example.

In another embodiment, the presence of a disease or disorder may be determined by sequencing one or more nucleic acid sequences of the invention and identifying a mutation or sequence aberration in a hypersensitivity site of a patient as compared to a normal control. In another embodiment, the presence of a disease or disorder may be determined based on differences in cleavage by a nuclease, chemical, or other agent used to detect hypersensitivity sites.

In a related embodiment, by comparing the hypersensitivity sites present in a diseased cell or tissue to those present in a normal or non-diseased cell or tissue, sites correlating (*e.g.* present or absent to a different degree) to the presence of any disease may be identified. Cells from a patient suspected of having a disease may then be examined or profiled to identify the presence or absence of one or more hypersensitivity sites associated with a disease or disorder. The presence or absence of a hypersensitivity site, as associated with a specific disease, indicates that the patient has the disease. In certain embodiments, hypersensitivity sites from a patient are determined and compared to databases or computer readable medium comprising sets of hypersensitivity sites associated with a disease to determine if the patient has a

disease. The patient's hypersensitivity site profile may be compared to profiles established for one or a plurality of diseases. Therefore, the method may be used to detect or diagnose disease in the absence of clinical symptoms or any other indication of the nature of the disease.

The invention also includes one or more sets of members (either regulatory sequences, genome locations or both) that are associated with genetic abnormalities of uncontrolled cell growth. A skilled artisan using the information may conveniently can prepare a set of genetic anomalies associated with one or more human diseases (*e.g.*, cancers, immune disorders, neurological disorders, cardiac disorders, or genetic disorders generally). For example, by matching known genetic changes associated with malignant transformation with the precise sequence or position information of a hypersensitive regulator, it is possible to identify genetic anomalies associated with a specific cancer. In one embodiment, regulatory sequences or units associated with a specific disease are identified by comparing hypersensitivity sites and/or their genome location identified in a disease sample as compared to those identified in a normal control sample. Identified regulatory sequences specifically associated with a disease may be used individually or as a set to detect or diagnose the disease or to identify regulated genes involved in disease onset or progression, for example. Accordingly, in one embodiment, the invention provides a method of detecting a disease or disorder, comprising identifying the one or more regulatory sequences or units in a subject and comparing these regulatory sequences to those of a normal control sample and/or a positive control sample. Sequences may be compared based on a variety of criteria, including, for example, their activity as determined by hypersensitivity assay, their chromatin structure (*e.g.* methylation or acetylation status), bound polypeptides, or sequence.

In a particularly useful embodiment a pre-existing set of genetic changes associated with a disease are compared with information from sets of genome locations of regulatory sequences to determine which of the changes linked to disease involve regulatory DNA. This information provides great value for drug discovery and for new modalities for treating disease.

In yet another particularly useful embodiment, allelic variants of regulatory DNA sequences are correlated with genetic diseases, drug treatments and responses thereto, effects of and responses to environmental or chemical exposures, or a variety of other outcomes. Allelic variants may be identified, for example, by comparing the

hypersensitivity at a specific site from samples derived from different individuals or cells, including individuals or cells with a disease or treated with a drug, for example. Such comparisons may be based upon alterations in accessibility or digestion at the site, ability to bind regulatory molecules, the presence or absence of chromatin modification such as methylation or acetylation, for example, chromatin or, alternatively, the sequence of a site.

The databases may be used to map disease causing SNPs (single nucleotide polymorphisms. Such single nucleotide polymorphisms, which cause changes to the expression pattern of nuclei, are more frequent within active genetic elements. A priori, the database of known regulatory sites may be screened to capture a population of phenotypically active SNPs.

The sequences may be used for toxicological profiling of potential drugs. Characterizing the molecular consequences of applying or titrating a drug into cell populations, tissues, or test organisms is very useful to define the pathways and side effects of a drug. Comparison of the patterns from hybridization experiments using the isolated regulatory sequences or units probed with the probes derived from the test populations can confirm the mechanism of the drug. Testing the response of patients to a regime of drugs also allows identification of patients who may be more or less suitable for that particular treatment, as some patients may show little induction of the target active genetic element or an unexpected activity in other sets of hypersensitive sites. Thus, in one embodiment, the invention provides a method of qualifying a patient for a clinical trial or for treatment with a drug or therapy that involves determining the hypersensitivity profile of the patient and comparing it to the hypersensitivity profiles of patients known to respond positively or negatively to a particular drug or therapy. Alternatively, the status of one or more individual hypersensitivity sites, regulatory sequences, or regulatory units may be used for such purposes.

In a related embodiment, the invention provides a method of correlating clinical data with hypersensitivity sites to predict the outcome of a disease or treatment protocol. Hypersensitivity site profiles may be established for patients and correlated with disease progression or outcome, alone or after treatment with any therapy or protocol. The hypersensitivity site profile may then be determined for a patient and used to predict disease outcome or the success of a given treatment protocol and will assist in determining the appropriate therapy. In addition, one or

more regulatory sequences or units associated with a particular clinical outcome or response to treatment may be identified, for example, by comparing hypersensitivity site profiles between patients with different responses and identifying one or more associated with a response. Clinical outcome may then be predicted by examining the activity or sequence of one or more of the identified regulatory sequences or sites in a candidate subject.

The sequences may also be used for discovery of novel lead compounds. Drug discovery can be advanced by understanding the biology of the target disease system and, in particular, the identification of key active genetic elements involved in disease progression. For example, high throughput screening using labeled probes able to detect the formation of hypersensitive sites in nuclei can be used to identify compounds that affect the formation of specific hypersensitive sites. Although the use of probes corresponding to hypersensitive sites is provided as an exemplary method, the activity of regulatory sequences and units associated with the hypersensitivity sites may be examined by any available means, including, for example, sequencing. This is true for all methods of the invention.

Hypersensitivity site profiles may be compared between cells treated with a drug and untreated or control cells to identify drugs and drug candidates. Furthermore, where a specific hypersensitivity site has been associated with a disease or disorder, probes specific for such site may be used to determine the status of the site before and after drug treatment to identify a drug that alters the status of the hypersensitivity site and would, therefore, be useful in therapy of the associated disease or disorder. In certain embodiments, treatment with the drug restores the status of the hypersensitivity site to that observed in a control sample. In one embodiment, the status of multiple regulatory sequences within a regulatory unit is examined to determine the effect of treatment with a drug or other agent.

Hypersensitivity sites and profiles thereof may also be used according to the invention for toxicology profiling of drugs. Determination of the effect of a drug upon hypersensitivity sites may be predictive of drug toxicology, for example, based upon the effects of known toxic agents or drugs upon one or more hypersensitivity sites. In another embodiment, drug toxicity may be correlated with specific patients based upon the presence or absence of one or more hypersensitivity sites, for example, those corresponding to regulatory sequences or units, in patients wherein a drug has been toxic. The ability to predict drug toxicity, particularly where only a relatively small

number of potential patients are susceptible, will allow physicians to selectively avoid treating patients potentially subject to drug toxicity. Thus, the invention provides a method of screening patients for drug treatment that includes examining the status of a regulatory sequence or unit of the invention and comparing its status to that observed in patients known to have a negative outcome upon drug treatment.

The invention further provides drugs identified according to any of the methods of the invention. For example, the invention provides drugs, including small molecules, for example, identified as affecting or altering the accessibility of on or more hypersensitivity sites, including, for example, hypersensitivity sites corresponding to a regulatory sequence or RS. Accordingly, the invention provides a drug produced by the process of screening one or more compounds for their ability to alter one or more hypersensitivity sites or the activity of one or more regulatory sequences or RSs, identifying a compound that alters one or more hypersensitivity sites or the activity of one or more regulatory sequences or RSs and producing said compound. Alterations in hypersensitivity sites may be detected based upon changes in their cleavage or accessibility by nucleases or other agents that cleave DNA, for example, and may involve either an increase or a decrease in hypersensitivity.

The invention further provides methods of manufacturing a drug of the invention. Such methods comprise identifying a drug that effects or alters one or more hypersensitivity sites by a method of the invention and producing the identified drug.

The invention provides therapeutic methods related to the control of gene expression. Such methods generally include increasing or decreasing expression of a gene, either directly or indirectly. As described *supra*, the invention provides methods of regulating the expression a gene using polynucleotides of the invention. In certain embodiments, the gene is a therapeutic molecule. For example, methods of the invention include providing a polynucleotide, for example, an expression vector, to a subject lacking adequate expression of a gene, wherein the polynucleotide or vector comprises a regulatory sequence or unit of the invention and the gene.

In certain embodiments, methods of the invention related to regulating gene expression include diminishing or disrupting expression of a target gene. Some regulatory sequences are known stimulate disease either through their formation or through overexpression of regulated genes. Accordingly, information regarding regulatory sequences and databases comprising the same, may be used for designing polynucleotides to interfere with the formation of regulatory sites. The sequence

information associated with such sites can be used to design molecules, such as polynucleotides, including knockdown reagents, as described *supra*, or synthetic chemicals, to block their formation in nuclei. Inhibition of the formation of some regulatory sites may be used to inhibit development of the development of the disease phenotype.

The databases may also be used for experimental control of transcriptional programming or gene expression. DNaseI regulatory sites had been shown to be responsible for positive and negative regulation of genes in nuclei. The databases of regulatory sites may be used to design molecules that can interfere with the formation of a functional regulatory site in the nuclei and so control transcriptional regulation. Such an experimental tool can perform functional gene knockout or knockdown experiments or otherwise examine the redundancy of the regulatory network in the eukaryotic nucleus. That is, the inhibition of the formation of a specific regulatory site may cause an expected alteration in the transcriptional program or induction of a different pattern of active genetic elements and, therefore, may be used to control expression of coordinately regulated genes, for example.

In one embodiment, a library is obtained from cells that have become genetically altered, such as tumor cells, and compared with the same but normal cell type. A comparison reveals regulatory systems involved (or lost) in the altered cells and that are relevant to exerting control over the cells. In a related embodiment, a cancer diagnosis and/or treatment, is provided comprising the steps of: providing a tissue biopsy; obtaining a library of regulatory sites from one or more cells of the biopsy; and comparing the sequence location, sequence, and/or abundance of regulatory sequences with another set of regulatory sequences. This method provides a set of regulatory sequences of great value for treatment of the disease. The "another set of regulatory sequences" in this context can include a set of information stored in a computer or other electronic medium, and may include a control set of sites obtained from the same patient or family member.

In a desirable embodiment, a further step is provided wherein the information obtained by the comparison is used to select an agent, such as a knockdown reagent, a specific inhibitor of protein-nucleic acid binding, a chemical agent known to interact with a specific protein linked to a specific regulatory event, a viral vector and the like. The selected agent is administered to the patient and alters regulation by up-regulating or down-regulating one or more particular regulatory mechanisms. Genome-wide

location analysis provides a powerful tool for further dissecting gene regulatory networks, annotating gene functions and exploring how genomes are replicated.

Other embodiments and uses of the invention will be apparent to those skilled in the art from consideration of the specification and practice of the invention disclosed herein. All references cited herein, including all U.S. and foreign patents and patent applications, are specifically and entirely incorporated by reference. In addition, U.S. patent applications No. 60/108,206, No. 09/432,576, No. 60/302,369, No. 60/290,036, No. 60/294,890, No. 60/294,890, No. 60/378,664, No. 60/387,910, No. 60/387,887, No. 10/187,887, and No. 60/404,121, and PCT applications PCT/US02/15032 and PCT/US02/16967 are specifically and entirely incorporated by reference. It is intended that the specification be considered exemplary only, with the true scope and spirit of the invention indicated by the claims.

6. EXAMPLES

The examples described below provide illustrative techniques according to certain embodiments of the present invention for characterizing regulatory sequences, for example those associated with a genetic locus of interest. It will be understood by the artisan of ordinary skill that the procedures and techniques described herein below are provided by way of illustration and not by way of limitation, and that the specific conditions and parameters described for these experiments can be varied while still achieving essentially the same or similar results and advantages according to the present invention.

6.1 Preparation of samples for use in genetic locus profiling

Nuclei from cultured primary cells, cell lines and primary tissues were prepared and treated with DNAse I as follow:

Tissue Prep Protocol

I. Buffer and Solution Preparation:

Unless otherwise noted, all buffers and reagents should be filtered (0.22μM) and pre-chilled to 4°C (on ice) before use.

Buffer A (per Litre):

Final Concentration	*Stock concentration*	*Amount used from stock*
15 mM Tris-Cl, pH 8.0	1 M Tris-Cl, ph 8.0	15 mL
15 mM NaCl	5 M NaCl	3 mL
60 mM KCl	3 M KCl	20 mL
1 mM EDTA, pH 8.0	0.5 M EDTA, pH 8.0	2 mL
0.5 mM EGTA, pH 8.0	100 mM EGTA, pH 8.0	5 mL
0.5 mM Spermidine	0.5 M Spermidine	1 mL
0.15 mM Spermine	0.5 M Spermine	0.3 mL

Combine appropriate amounts of stock solutions and add sterile dH_2O to a final volume of 1 Litre.

10% NP40 (per 100 mL):

Final concentration	Stock concentration	Amount used from stock
10% NP40	100% NP40	10 mL

Combine 10 mL 100% NP40* and 90 mL nuclease free, sterile dH_2O.

10% NP40 should be warmed in a 55 °C waterbath for 30 minutes prior to use to ensure proper dissolution.
Store at 4 °C

5 M Urea, 2 M NaCl (per Litre)

Final concentration	Stock concentration	Amount used from stock
5 M Urea	Dry crystal (ultrapure)	300 g
2 M NaCl	5 M NaCl	400 mL

Combine 5 M NaCl solution and 300 g Urea. Fill to 800 mL with sterile dH_20 and stir until thoroughly dissolved ($\leq$30 minutes). After solutes have dissolved, fill to 1 Litre with sterile dH_2O and stir briefly to thoroughly mix. Store at room temperature.

DNaseI 10X Digestion Buffer (per 100 mL)

Final concentration	Stock concentration	Amount used from stock
60 mM $CaCl_2$	1 M $CaCl_2$	6 mL
750 mM NaCl	5 M NaCl	15 mL

Combine stock solutions, and 79 mL nuclease free, sterile dH_2O. Store 1 week only at 4 °C.

Stop Buffer (per Litre)

Final concentration	Stock concentration	Amount used from stock
50 mM Tris-Cl, pH 8.0	1 M Tris-Cl, pH 8.0	50 mL
100 mM NaCl	5 M NaCl	20 mL
0.10 % SDS	20% SDS	5 mL
100 mM EDTA, pH 8.0	0.5 M EDTA, pH 8.0	200 mL

Combine stock solutions and add sterile dH_2O to a final volume of 1 Litre. Dispense into 50-mL aliquots and add 1:1000 volume of RNase (from 10 mg/mL stock) (Roche Applied Science, Indianapolis, IN). For example, for each 50-mL aliquot of Stop

Buffer, 50 µL of enzyme stock should be added. Store at 4 °C until ready for use. Immediately prior to use, heat stop buffer aliquots to 42 °C in waterbath.

II. Tissue Preparation

The protocols below describe how the procedures were performed.

A. Tissue Collection

1. Obtain fresh tissue and maintain on ice. Tissue should not be frozen. Handle at 4°C and process tissue as quickly as possible. The following protocol is designed for 0.2g-2g of tissue. Tissue samples with mass greater than 2g should be cut into pieces that are <2g and processed separately.
2. Record the arrival time and harvest time of the tissue. Spin the tissue samples at 500 xg for 3min in a 4°C centrifuge. While tissue is spinning, remove buffer A, 10x DNAse digestion buffer and 10% NP-40 from 4°C storage and keep on ice.
3. After spin, decant the media and weigh the individual samples using one of the receiving tubes as a tare. Record the weights and transfer the tissue to a 50-ml Falcon tube on ice.
4. Add 25ml of buffer A and homogenize for 30 seconds using a PowerGen 125 at max speed. Take a small aliquot out of each tube, dilute in trypan-blue and view under microscope for a viability assessment. Spin tubes at 500 xg for 3min. The tubes and centrifuge should be kept at 4°C for entire procedure.
5. Spin at 500 xg for 3 min and pass sample though the MACS Cell Isolation columns (Miltenyi Biotech GmbH, Germany). Follow the MACS protocol and resuspend the cells in 25ml of Buffer A. Take a cell count of viable cells using trypan-blue. (10ul of cells to 40ul trypan blue, 1/5 dilution)

6. Wash cell pellet twice with 25ml of buffer A (500 xg for 3 min). Resuspend cell pellet by pipetting up and down and spin at 500 xg for 3 min at 4°C. At this point the cell pellet should be white or white/yellow. If the cell pellet is not white/yellow then additional washes should be performed.
7. Resuspend pellet in 25ml of buffer A.

B. Nuclei Isolation

1. Using a cut 1000-ul pipette tip (angled cut about 3-5mm from tip) add 1ml of ice cold 10% NP-40 drop wise. It is important that the NP-40 be added slowly one drop at a time with constant swirling. When the entire volume/1 ml had been added, mix well and incubate on ice for 10min.
2. Spin nuclei at 500 xg for 3min. Carefully pour off buffer A and resuspend/wash with 25ml of fresh buffer A. This is to remove all of the NP-40. (NP-40 is a detergent that will lyse the nuclei given enough time, so removing all of it promptly is important.)
3. Collect nuclei by centrifuging at 500 xg for 3min. Resuspend nuclei in a volume that will give a concentration of approximately 100 million nuclei per ml based on the cell count from step 5 above.

C. Nuclei Count

1. Make a series dilution (100ul nuclei product in 900µl buffer A, x2) to give a 1/100 dilution product.
2. Count 4 separate cells (squares) on the Hemacytometer and take the average of the 4 counts to give you an accurate nuclei count/ml. (multiply the average number of nuclei per 1 square by the dilution factor (100) and 10^4 to calculate the nuclei per ml)

D. Time Trial Digestion

3. Add 1/1000 volume of both Proteinase K (Roche Applied Science, Indianapolis, IN) and RNAse to stop buffer, and warm up in 42°C water bath.
4. Calculate total nuclei yield (nuclei count/ml x total volume). The total number of 1ml reactions that can be done is determined as follows:

reactions = nuclei yield/$20x10^6$(nuclei/reaction)

That is, each digest contains $20x10^6$ nuclei. Next calculate the product volume that will give 20 million nuclei:

Product volume(ml/reaction) = $20x10^6$(nuclei/reaction)/nuclei count(nuclei/ml).

5. Take the product volume calculated in step 4 and add that volume to one or two tubes, then add the appropriate amounts of 10x DNAse Buffer 100ul per 1ml reaction and Buffer A:

Volume Buffer A = 1ml - (product volume+10x DNAse Buffer volume)

Set these tubes aside; they will be the untreated or 0 samples.

6. In another tube add all the Buffer A, 10x DNAse Buffer and the appropriate amount of DNAse I enzyme (Roche Applied Science, Indianapolis, IN) for all the remaining tubes calculated for in step 4 (DNAse stocks are kept in aliquots labeled with units/µl, and reactions are carried out with units/100 µl reaction volume). Place this tube in a 42°C water bath.

7. Label tubes with cell type, DNAse concentration and reaction time (K562-2-15sec). Label as many as were calculated for in step 4 minus 1 (the minus 1 accounts for any pipetting error). Place these tubes in 37°C water bath and add 1ml of stop buffer from step 3.

8. Preheat at 37°C the zero tubes from step 5 and the remaining nuclei product tube. The zero tubes only need 1 minute then add 1ml of stop buffer from step 3. Incubate the nuclei product tube should incubate for 2.5 minutes unless the total volume is greater than 5ml. If the volume is greater than 5ml use the fluke thermometer and incubate until the product temp is 30°C.

9. Mix the preheated nuclei from step 8 with the reaction solution from step 6. This begins the DNAse reaction. Using a 1000µl tip with the end angled off (as in step 1 of Nuclei Isolation) take 1ml aliquots out at the appropriate time and dispense into the correctly labeled tubes from step 7. Thus stopping the DNAse reaction at various time points.

10. When the DNAse reaction is complete, the tubes are transferred to a 55°C waterbath and incubated for 15 minutes to allow for RNA digestion by the Rnase in the stop buffer. 1:1000 volume of proteinase K (from 20mg/mL stock) is added to each reaction tube (e.g. 10µL proteinase K for 10mL reaction with 10mL stop buffer) and left overnight to allow for protein digestion.

6.2 Purification, preparation and treatment of DNA for use in genetic loci profiling

A. DNA purification and quantitation

After digestion of nuclei with Proteinase K the DNA was purified using the Puregene system (Gentra, Minneapolis, MN) according to the protocol titled "DNA Purification, PureGene protocol" and quantitated by UV spectrophotometry.

PureGene protocol:

Materials

Cell lysis solution

Protein precipitation solution

100% isopropanol

70% Ethanol

10 mM Tris-Cl, pH 8.0

15 or 50-ml Falcon tubes.

55 °C waterbath

Vortex

This protocol is scalable, and based on an initial concentration of 6-7 X 10^6 nuclei/ml. If the initial concentration is too high, add cell lysis solution to reach a final concentration of 6-7 X 10^6 nuclei/ml.

Protocol

1. To nuclei prep (treated with both RNase and proteinase K, and incubated overnight at 55°C) add protein precipitation solution to a final concentration of 25% v/v. Vortex for 20 seconds and incubate on ice 5 min.
2. Centrifuge at 3200 X *g* for 10 min. at room temperature to pellet any precipitated materials. If a pellet does not form or was too loose, repeat steps 1 and 2 before continuing on to 3.
3. Pour the supernatant (containing DNA) from step 2 into a new, labeled falcon tube containing 0.7 X volume of 100% isopropanol. Invert ~50 times and incubate at room temperature 5 min.
4. Centrifuge at 3200 x *g* for 10 minutes to pellet all precipitated DNA. Carefully pour off supernatant without disturbing the DNA pellet.
5. Wash the DNA pellet with 2X starting volume 70% ethanol, invert several times, and centrifuge at 3000 x *g* for 5min.
6. Carefully pour off the supernatant, taking care not to dislodge the pellet. Gently invert the tube on an absorbent pad for 10-15 min. to dry the pellet. The pellet may be additionally dried using the speedvac.

7. Resuspend the DNA pellet in 10 mM Tris-Cl, pH 8.0 in a volume of about 1-ml per 20 X 10^6 starting nuclei.
8. Incubate at 55 °C for one hour (Note: depending upon the need for high molecular weight DNA, the resuspended DNA may be sheared by syringe or hydroshear to aid in a more even resuspension).

B. Terminal transferase

Samples that had been treated with > 4U (/100μL of nuclei) of DNase I were treated with terminal transferase to add a di-deoxy nucleotide to the ends of all DNA fragments. This procedure inhibited the formation of an artifact during the initial cycles of the qPCR reactions by preventing extension of small DNA fragments, generated during the DNase I treatment. DNA samples were treated under the following conditions: 1X Tdt buffer, 2.5 mM $CoCl_2$, 0.5mM ddNTP blend and 0.005U/μL terminal transferase (Roche Applied Science, Indianapolis, IN). Reactions were incubated at 37C overnight and re-purified using the Puregene system (Gentra Systems, Minneapolis, MN).

C. Shearing

Both calibrator and "treated" DNA samples were mechanically sheared using a HydroShear (GeneMachines, San Carlos, CA) to uniformly produce fragments of 5-10kb in length. This was done to insure complete melting of the template DNA during the initial cycles of the PCR. Genomic regions difficult to melt may yield erroneous results due to underestimation of initial starting copy number. After shearing, DNAs were quantitated by UV spectrophotometry and adjusted to ~ 10ng/μl.

D. Quality control of DNAs used for locus profiling

Prior to using a particular DNA for locus profiling, the sample was run against a number of control loci. For example, HS2 (hypersensitive site 2) of the human beta-globin LCR was routinely analyzed in the cell line, K562 and fetal liver. The HS score at this site allowed us to determine the degree to which a DNase I digestion proceeded. A sample that had been over-digested with DNase I will result in "noisy" locus profile, making it difficult to identify new HS sites. In contrast, a sample that

was under-digested will not result in substantial copy loss, reducing the sensitivity of the assay. This was an important step before proceeding with a profiling experiment.

6.3 Design and preparation of primers for use in genetic locis profiling

In one illustrative embodiment, in order to identify the DNA hypersensitivity sites associated with a genetic locus of interest, oligonucleotide primers were designed that were capable of amplifying products, for example overlapping products, that span substantially the entirety of the genetic locus. In this example, the software program Primer3 v.2[1] (Rozen S and Skaletsky HJ. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386) was employed using the following parameters:

Primer Tm: 60C optimal (+/- 2C)

1. Primer length: 22 bases optimal (ranging from 18-25 bases)
2. Product size: 250 bases optimal (ranging from 225 to 275 bases)
3. Percent GC: 50% optimal (ranging from 40 to 80%)
4. Max PolyX: 4
5. All other settings: default

In designing a set of primers for a genetic locus profiling experiment, one strategy was to cover the genomic region as completely as possible by walking stepwise, 5' to 3', in approximately 250 bp fragments. Once a region was selected for primer design the sequence was preferably repeat masked. A design window of 300 bp was selected at the starting point. Once the first amplicon in designed, a second window was selected 20 bp before the 3' end of the first window. Each primer will be allowed a maximum of 10 matches of the 16 bases on the 3' end with the genome. Occasionally, Primer3 will not be able to design primers within a given window. In such a case, the window may be shifted 5 bp in the 3' direction and another attempt made to design and identify suitable primers. This occurs until a successful primer pair was chosen. This process continues in a sequential manner until the 3' end of the locus had been reached.

It was desirable to have all primer pairs amplify with the *same* master mix and the *same* set of thermalcycling conditions as this allows the interrogation of many different amplicons simultaneously. It was more difficult and labor-intensive to perform locus profiling if every pair of primers had to be optimized separately and

required specific amplification conditions. Because the genomic targets being amplified can vary extensively in their GC content and can sometimes be difficult to amplify with high efficiency, the primer design plays an important role in effective locus profiling.

By increasing the optimal length to, for example, 24+ bases, it may be easier to manipulate the annealing temperature. Elevating the annealing temperature during thermalcycling and increasing primer length will yield a higher degree of binding specificity, thereby reducing the number of amplicons that yielding multiple amplification products. Moreover, adjusting the master mix formulation in conjunction with increasing primer length, as described above, will result in more robust amplification and fewer primer pair failures. One illustrative formulation in this regard includes three agents (glycerol, BSA and betaine) to improve amplification of targets high in GC content.

6.4 Quantitative PCR (qPCR) method for genetic locus profiling

A. Dilution and mixing of primers

Primers were received in 96-well microtiter plates from the manufacturer (Illumina, San Diego, CA) in "left" and "right" plates. These primers were supplied at normalized 50μM stocks. Prior to setting up qPCR reactions, a working, mixed dilution (1μM) was prepared using a liquid handling robot (Biomek FX, Beckman, Fullerton, CA). Using a single 96-well transfer protocol, 10μL of stock primer was removed from each well of the "left" plate and added to a 96 deep-well plate containing 480μL of PCR-grade water in each well. This was then repeated for the "right" plate, adding it to the 96 deep-well plate. The plate was sealed and shaken to mix the primers. This plate was then referred to as the "Primer Pair Plate" or PPP.

B. Rearraying of primers from Primer Pair Plates

To assemble the individual qPCR reactions, the appropriate amount of primer mix was dispensed into a new 96-well PCR plate (18μL of 1μM mix/well), where it will eventually be combined with the DNA template and master mix. The layout of such a plate is shown in Figure 12a.

C. Combination of template DNAs with qPCR master mix

Pre-master mix was removed from the –20°C freezer, thawed, and water and FastStart Taq polymerase (Roche Applied Science, Indianapolis, IN) were added. Upon addition of water and polymerase this mix was referred to as qPCR master mix. A detailed protocol for the assembly of an illustrative qPCR master mix was as follows:

qPCR master mix was composed of 1X FastStart buffer (Roche Applied Science, Indianapolis, IN), 3mM $MgCl_2$, 200μM of each dATP, dCTP, dGTP and dTTP, 0.8% glycerol, 0.5M betaine, 0.5mg/mL BSA, 300nM 6-ROX (Molecular Probes, Eugene, OR), 0.33X SYBR Green I nucleic acid stain (Molecular Probes, Eugene, OR) and 0.033U/μL FastStart Taq polymerase (Roche Applied Science, Indianapolis, IN).

	100 mLs qPCR pre-master mix (2X)		
Step 1			
		MLs	
1	10 mM dNTP mix	4.2	
2	25 mM MgCl2	25.2	
3	100 mg/ml BSA	1.05	
4	50% glycerol	3.36	
5	5M Betaine	21	
	Mix together reagents 1-4 and aliquot 10.440 mLs into each of 5 50 mL conicals		
Step 2			
		MLs	
5	10X FastStart buffer	20.5	Buffer w/o MgCl2
6	6-ROX	0.05	
7	SYBR Green I	0.006	
	Add 50 ul of 6-ROX (shake tube before taking aliquot from surface of liquid) to the 10X FastStart buffer, add 6 ul of SYBR green I to that combination.		
	Make sure SYBR green I is completely thawed and mixed.		
	Add 4.011 mLs of this mixture to the 10.440 mLs added in Step 1.		
	Mix thoroughly and freeze at -20C. These dyes are photosensitive.		
Step 3			
	Each conical will yield 20 mLs of 2X master mix.		
	Thaw, add 5.283 mLs of PCR grade H2O and 266 ul of FastStart Taq		
	Mix thoroughly and protect from light.		

1. Assembly of the DNA master mix plate:

Calibrator and treated DNAs were mixed with 2X qPCR master mix in 50 mL conical tubes and then aliquoted into a 96 deep-well plate; column 1 containing "calibrator" master mix and column 2 containing "treated" master mix, alternating across the entire plate (e.g., see Figure 10). The amount of each DNA/master mix prepared depends on how many qPCR reaction plates were being assembled at a time. Details regarding illustrative volumes used per plate were as follows:

qPCR DNA/Master Mix Set up:

Number of Plates Per Run	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Calibrator DNA	1380	2070	2760	3450	4140	4830	5520	6210	6900	7590
Treated DNA	1380	2070	2760	3450	4140	4830	5520	6210	6900	7590
Master Mix for Calibrator DNA	3450	5175	6900	8625	10350	12075	13800	15525	17250	18975
Master Mix for Treated DNA	3450	5175	6900	8625	10350	12075	13800	15525	17250	18975
Dispense (ul) into each well:	92	137	182	227	272	317	362	407	452	497

2. Assembly of qPCR reference plate:

The qPCR reference plate provides the standard curves and thereference amplicons from which all copy number comparisons were made. Three reference primer pairs (referred to as PPT-1, PPT-2 and PPT-3 on Figure 11), also at a concentration of 1μM, were combined with calibrator master mix, treated master mix and master mixes created for each of 4 standard DNAs. These master mixes were then dispensed into a 96 deep well plate. Details regarding illustrative volumes used per plate were as follows:

qPCR reference plate setup:

Volumes for each standard DNA for the number of plates being run

Number of Plates Per Run	1	2	3	4	5	6	7	8	9	10
1 uM Primer Mix (uL)	18	36	54	72	90	108	126	144	162	180
Standard DNA (uL)	12	24	36	48	60	72	84	96	108	120
Master Mix (uL)	30	60	90	120	150	180	210	240	270	300
Dispense into each well (uL)	20	40	60	80	100	120	140	160	180	200

D. Creation of the qPCR reaction plate

Before 15μL qPCR reactions can be dispensed onto a 384-well reaction plate, the qPCR DNA master mix plate was combined with the rearrayed primer plate (Figure 12a and 12b). After the addition of 42μL of the DNA/master mix to each well of the rearrayed primer plate, the plate (now referred to as a DRAP) was heat-sealed and mixed by inverting the plate and agitating on an orbital shaker. This plate was then returned to the upright position and the mixture centrifuged to the bottom of the wells. At this point, all of the components of the qPCR reaction had been assembled in a large aliquot which was further subdivided into 3, 15μL reactions on a 384-well plate.

Triplicate, 15μL reactions were transferred (via 96-well, robotic manipulations) into quadrants 1, 2 and 3 of an ABI, 384-well reaction plate. A single, 96-well transfer of 15μL was made from the qPCR reference plate into the 4th quadrant of the same ABI, 384-well reaction plate. The 384-well plate was then sealed with optical tape and centrifuged at 4000 RPM in an Eppendorf refrigerated centrifuge for 1 minute. The reaction plate was then placed on a 7900HT ABI Prism Sequence Detection System for thermalcycling.

E. Thermalcycling of qPCR reactions

qPCR reactions were denatured at 95C for 10', followed by 6 cycles of 98°C, 30s; 55°C, 30s; 72°C, 60s; followed by 29 cycles of 95°C, 30s; 55°C, 30s; 72°C, 60s. After the final cycle, a melting curve was performed by slowly increasing the temperature (2 % ramp) from 65°C to 95°C and continuous fluorescence acquisition. All reactions were performed using a 7900HT ABI Prism Sequence Detection System (Applied Biosystems, Foster City, CA).

F. Data Analysis

After the reactions were complete, the normalized fluorescence data was exported using the ABI SDS software and then analyzed using software that determines the starting copy number for every target amplicon relative to each of the reference amplicons. A detailed description of an illustrative analysis algorithm was provided below in Example 6.5.

Because SYBR green I was used to detect the accumulation of PCR product during amplification it was necessary to examine the melting curve for each set of reaction replicates. SYBR green I will detect any product generated during the PCR reaction. A reaction that yields multiple products or primer dimers will yield erroneous results because the fluorescence of all double-stranded DNA was acquired by the 7900HT SDS. Different DNA fragments will typically possess unique melting behaviors, thereby allowing a more rapid determination of which reactions contain more than one product. It was preferred in this assay that only one product was amplified.

G. Analysis of Locus Profiles

The goal of performing a locus profile was to identify regulatory sequences, for example DNase I hypersensitive sites, within and surrounding a genic region of interest. Locus profiles were typically designed to cover the promoter, the first few introns and immediately 3' of the last exon of a gene under the presupposition that these segments were the most likely to contain regulatory elements and exert control over gene expression. The size of a locus profiling experiment depends, in part, on the size of the gene or gene cluster being analyzed. A locus profile comprised of 250 bp amplicons typically spans 20-50 kb in size. By performing quantitative PCR with each amplicon, a loss in copy number, resulting from digestion with DNase I can be reproducibly detected.

After the HS score was calculated for each amplicon by the analysis tool, the values were plotted vs. the absolute genomic position to yield a DNase hypersensitivity graph.

In one illustrative example of how locus profiling data may be represented, the HS score was determined in a relative fashion by comparing the copy number of each target amplicon to a reference. Reference amplicons had been selected from genes

which were not expressed at an appreciable level in the cell typed being examined by the locus profiling experiment. When a gene was not expressed, the chromatin was believed to be in a closed conformation. In such a case, DNase I does not have easy access to the DNA and cannot digest it when it was wound around nucleosomes. The reference amplicon allows us to estimate the copy number of a DNA sample at a site that was not susceptible to DNAse I digestion.

A HS score of 1 indicates equal copy numbers of the reference amplicon and target, showing no copy loss due to digestion with DNase I. In general, HS scores of 0.75 or less can be designated as DNAse I sensitive. In the example below, two different DNAse I treated samples had been profiled; a 4U and an 8U sample. Typically, the HS score of a DNase I sensitive site was higher (indicating less sensitivity) in a sample that had been treated with less enzyme and lower (indicating more sensitivity) in a sample digested with more. When superimposing the graph of the 8U-treated sample over the 4U, one can see that true HS sites become more sensitive as the amount of DNase I was increased (Figure 13a).

The fact that multiple, sequential amplicons display copy loss was a good indicator DNase I hypersensitivity. The boxes in Figure 13b highlight the HS sites within the profile. It was possible that multiple, discrete HS sites were located within the 250 bp fragment surveyed. To further elucidate HS sites, it may be possible to re-profile HS regions with smaller or overlapping amplicons resulting in a profile with higher resolution.

Figures 13a and 13b show the conventional representation of locus profiling data; an alternative method in which the –log of the HS score, multiplied by 100 can give another view of the data seen in Figure 13c below. In this case, the HS scores were higher numbers and stand out more dramatically from the background.

To validate the HS sites identified by locus profiling, qPCR reactions were repeated in multiple samples of the same cell type from which the sites were originally identified.

6.5 Illustrative qPCR data analysis algorithim

The throughput of gene quantification using Quantitative PCR (qPCR) in a production environment was typically constrained by two parameters; the cost of the material (enzymes, reagents, etc) used in the experiment and the number of samples that can be processed in a given timeframe. Current methodologies require the

determination of the amplification efficiency of the RTPCR event using a "standard curve" technique. In this technique a serial dilution of a known quantity of DNA/RNA was typically used to calibrate the sample of unknown quantity. The dilutions generate the "standard curve".

Benefits of the Analysis Technique

1) Increased throughput

a) The conventional operational approach to qPCR involves the determination of a standard curve for every amplification product to determine the amplification efficiency of the primer pair. In typical experiments nine to twelve experimental measurements were required for each primer pair. In the high throughput environment using 384 well reaction plates the 24 individuals pairs can be effectively measured using the conventional approach.

b) The approach described in this treatment increases the throughput to 48 individual primer pairs for a 384 well reaction plate without affecting the accuracy or precision of the experiments.

2) Absolute copy number

Conventional approaches cannot easily correlate the initial copy number of amplifiable sequences appropriate for a given primer pair. This treatment outlines a technique for comparing the absolute copy loss of an ensemble distribution based on the bulk concentration of the DNA sample. (Cases 3 and 4)

This technique allows for a direct comparison of copy number changes between tissues and cell lines as a function of DNAse treatment.

DNAse Kinetics

In the simplest form, DNAse kinetics can be modeled as a pseudo-first order reaction when the protein concentration was in vast excess over the DNA for a bimolecular process:

$$P + DNA \xrightarrow{k_n} P - DNA$$

Where k_a was the bimolecular association constant (and k_d would be the bimolecular dissociation constant). Under pseudo-first order concentrations the system yields an observed association rate, k_{obs}. The latter can be used in fitting the time dependent data to the following equation:

$$Y = 1 - \exp(-k_{obs}t)$$

Of course, this model not only assumes that DNAse concentration was in vast excess, but it also assume that the system was homogeneous and that the fluctuations in the ensemble do not exist on the timescale of the process. This approximation was not appropriate in a biological process involving native chromatin in the nucleus. By applying the Fokker-Planck formalism it was possible to model the stochastic behavior of the heterogeneous system.

Theoretical Model:

In this derivation certain assumptions were made. First, the number of fluorescent molecules contributing to the RT-PCR signal was proportional to the length of amplicon, which was amplified. Second, there was only one product contributing to the fluorescence. And, third, the amplification efficiency was constant in the low fluorescence region.

In this treatment, one was calculating the hypersensitivity, σ, which should vary between 0 and 1. As the value of σ was a direct measure of the level DNAse cutting relative to a reference site.

Case 1: Ideal Model:

Initial condition:

Let L_1N_0 = number of bases in the initial sample with no DNAse treatment, and

L_1N_x = number of bases in the initial sample that had been DNAse treated.

Here, N_0 was the number of amplicon copies in the untreated sample, N_x was the number of amplicon copies in the treated sample, and L_1 was the length of amplicon. The number of bases at any point of the amplification can be expressed using the following relationships.

$$L_1 N_{n_a} = (L_1 N_0)\beta_1^{n_a} \quad (1)$$

$$L_2 N_{n_b} = (L_2 N_0)\beta_2^{n_b} \quad (2)$$

$$L_1 N_{n_c} = (L_1 N_x)\beta_1^{n_c} \quad (3)$$

$$L_2 N_{n_d} = \sigma(L_2 N_x)\beta_2^{n_d} \quad (4)$$

Where β_1 and β_2 were the PCR efficiencies of the reference amplicon and the target sample amplicon, respectively, and σ was the hypersensitivity. 'n' was the cycle number. In equations (1)-(4) the fluorescence was expressed in terms of the length of the amplicon and the number of copies. Explicit in the relationships was the assumption that the magnitude of the interaction of the fluorophores with the amplicon was proportional to the size of the amplicon. Implicit was the assumption that the fluorophore interaction was not strongly dependent on amplicon sequence.

Equation (1) – (4) can be expressed as:

$$\ln(L_1 N_{n_a}) = \ln(L_1 N_0) + n_a \ln \beta_1 \quad (5)$$

$$\ln(L_2 N_{n_a}) = \ln(L_2 N_0) + n_b \ln \beta_2 \quad (6)$$

$$\ln(L_1 N_{n_c}) = \ln(L_1 N_x) + n_c \ln \beta_1 \quad (7)$$

$$\ln(L_2 N_{n_d}) = \ln(\sigma L_2 N_x) + n_d \ln \beta_2 \quad (8)$$

At any set fluorescence value, the following relationships were true for the four samples.

$$\ln(L_1 N_{n_a}) \equiv \ln(L_2 N_{n_b}) \equiv \ln(L_1 N_{n_c}) \equiv \ln(L_2 N_{n_d}) \qquad (9)$$

If the fluorescence can be precisely determined for the four samples, equation (9) can be leveraged. From equations (5) and (6);

$$\ln(L_1 N_0) + n_a \ln \beta_1 = \ln(L_2 N_0) + n_b \ln \beta_2 \qquad (10)$$

Which simplifies to

$$\ln\left(L_1 / L_2\right) = n_b \ln \beta_2 - n_a \ln \beta_1 \qquad (11)$$

Also, from (7) and (8);

$$\ln(L_1 N_x) + n_c \ln \beta_1 = \ln(\sigma L_2 N_x) + n_d \ln \beta_2 \qquad (12)$$

That simplifies to

$$\ln\left(L_1 / L_2\right) - \ln \sigma = n_d \ln \beta_2 - n_c \ln \beta_1 \qquad (13)$$

From equation (11);

$$\ln \beta_1 = 1/n_a \left[n_b \ln \beta_2 - \ln\left(L_1 / L_2\right) \right] \qquad (14)$$

Combining equations (13) and (14) yields

$$\ln\sigma = \ln\left(L_1/L_2\right) + n_c/n_a\left[n_b \ln\beta_2 - \ln\left(L_1/L_2\right)\right] - n_d \ln\beta_2 \quad (15)$$

Or

$$\ln\sigma = \ln\left(L_1/L_2\right) + \frac{n_c}{n_a}\left[\left(n_b - \frac{n_d n_a}{n_c}\right)\ln\beta_2 - \ln\left(L_1/L_2\right)\right] \quad (16)$$

It was immediately obvious from equation (16) that only one PCR efficiency coefficient was necessary to determine the hypersensitivity of a sample. Specifically, if the coefficient was known for a reference amplicon, any sample can be evaluated with respect to the PCR efficiency of the reference amplicon.

Alternatively, one can solve for the hypersensitivity using amplification efficiency of the reference amplicon, which was shown in equation (17).

$$\ln\sigma = \ln\left(L_1/L_2\right) + \frac{n_d}{n_b}\left[\left(\frac{n_c n_b}{n_d} - n_a\right)\ln\beta_1 - \ln\left(L_1/L_2\right)\right] \quad (17)$$

Case 2: Non Ideal Copy Numbers

Initial conditions:

$$L_1 N_{n_a} = (L_1 N_0)\beta_1^{n_a} \quad (18)$$

$$L_2 N_{n_b} = (\eta L_2 N_0)\beta_2^{n_b} \quad (19)$$

$$L_1 N_{n_c} = (L_1 N_x)\beta_1^{n_c} \quad (20)$$

$$L_2 N_{n_d} = \sigma(\eta L_2 N_x)\beta_2^{n_d} \quad (21)$$

Here, the initial concentration of the reference amplicon and the target amplicon were not equivalent. This was possible, for example, when the PCR primers bind to multiple sites in the genome that can be amplified or the target amplicon was polyploid.

Using the same fluorescence argument as in Case 1, we arrive at equation 22.

$$\ln\sigma = \ln\left(L_1/L_2\right) + \left(\frac{n_d}{n_b} - 1\right)\ln\eta + \frac{n_d}{n_b}\left[\left(\frac{n_c n_b}{n_d} - n_a\right)\ln\beta_1 - \ln\left(L_1/L_2\right)\right] \qquad (22)$$

As η approach 1, equation 22 reduces to 17. Interestingly, for large η

$$\sigma \to \eta^{\left(\frac{n_d}{n_b}-1\right)}$$

Case 3: Calculating the Absolute Copy Number Loss

The number of bases at any point of the amplification can be expressed using the following relationships.

$$L_1 N_{n_a} = (L_1 N_0)\beta_1^{n_a} \qquad (23)$$

$$L_2 N_{n_b} = (L_2 N_0)\beta_2^{n_b} \qquad (24)$$

$$L_2 N_{n_d} = \sigma(L_2 N_x)\beta_2^{n_d} \qquad (25)$$

Here, by leveraging the relationship expressed in equation (9), one can, by combining (24) and (25), arrive at:

$$\ln(L_2 N_0) + n_b \ln\beta_2 = \ln(L_2 N_x \sigma) + n_d \ln\beta_2 \qquad (26)$$

And by combining (23) and (24), one can arrive at:

$$\ln(L_2 N_0) + n_b \ln\beta_1 = \ln(L_1 N_0) + n_a \ln\beta_1 \qquad (27)$$

This leads to the following result:

$$\sigma = \exp\left\{\ln\left(\frac{N_0}{N_x}\right) + \frac{(n_b - n_d)}{n_b}\left[\ln(L_1 N_0) - \ln(L_2 N_0) + n_a \ln \beta_1\right]\right\} \qquad (28)$$

The hypersensitivity can be expressed as:

$$\sigma \equiv \frac{N_x - N_c(\varepsilon)}{N_x} \qquad (29)$$

Where N_x was the average number of copies in the treated sample based on the bulk absorbance and $N_c(\varepsilon)$ was the number of copies cut at the DNAse concentration ε. Alternatively:

$$\frac{N_c(\varepsilon)}{N_x} = 1 - \exp\left\{\ln\left(\frac{N_0}{N_x}\right) + \frac{(n_b - n_d)}{n_b}\left[\ln(L_1 N_0) - \ln(L_2 N_0) + n_a \ln \beta_1\right]\right\} \qquad (30)$$

Case 4: Calculating Absolute Copy number with Repetitive Elements

Let $\Re$ represent the number of repetitive elements of the sample amplicon present in the genome G. In this case $\Re$ was determined relative to the number of repetitive elements of the reference amplicon.

$$L_1 N_{n_a} = (L_1 N_0)\beta_1^{n_a} \qquad (31)$$

$$L_2 N_{n_b} = (L_2 N_0 \Re)\beta_2^{n_b} \qquad (32)$$

$$L_2 N_{n_d} = \sigma(L_2 N_x \Re)\beta_2^{n_d} \qquad (33)$$

It can easily be shown that:

$$\sigma = \exp\left\{\ln\left(\frac{N_0}{N_x}\right) + \frac{(n_b - n_d)}{n_b}\left[\ln(L_1 N_0) - \ln(L_2 N_0 \Re) + n_a \ln \beta_1\right]\right\} \qquad (34)$$

Or

$$\frac{N_c(\varepsilon)}{N_x} = 1 - \exp\left\{\ln\left(\frac{N_0}{N_x}\right) + \frac{(n_b - n_d)}{n_b}\left[\ln(L_1 N_0) - \ln(L_2 N_0 \Re) + n_a \ln \beta_1\right]\right\} \quad (35)$$

It was clear that the repetitive elements in the system have a relatively weak dependence of the calculation of the cutting ratio.

It was apparent that the evaluation of the ratio was more strongly dependent on an accurate determination of the copy number of the bulk treated sample.

Data Manipulation

An illustrative program written in Java can be used to calculate the hypersensitivity, σ. The following block diagram depicts the major logical portions of the program.

Application Methodology

There are two components to the utilization of equation (16). First, it was necessary to determine β for the reference amplicon. Second, to apply equation (16) to RT-PCR results a high throughput mode, it was important to quantitatively determine a fluorescence intensity value for which the model was valid.

Determination of the PCR Efficiency Coefficient - β

There are two ways to determine β. The first technique relies on an experimental measurement of β by diluting a known copy number of the reference amplicon and measuring the cycle number ('*n*') at a constant fluorescence. β was simply the slope of '*n*' versus the initial copy number. This approach, while effective, does not fully utilize all of the information inherent in the experimental data. In addition, it requires the experimental determination of β for every plate in a throughput mode. At a minimum, a reasonable evaluation of β requires at least nine wells on the plate to

effectively determine the coefficient. The approach explicitly assumes that only one type of component (amplicon) was contributing to the fluorescence.

An alternative, and in some instances preferred, approach determines β directly from the reference amplicon curves directly. At low amplicon copy number, there are no limiting reagents in the sample. For example, relative to the amplicon copy number there are infinitely many primers and bases in the solution. In this limit PCR amplification was by far the major chemical event. Therefore, the following equation should represent the total fluorescence intensity from a sample at low copy number

$$R(n)=\sum_{i=1}^{\infty}(L_i N_i)\beta_i^n \quad (17)$$

Where $R(n)$ was the total fluorescence intensity at 'n'. From equation (17) we see that the fluorescence was now expressed as a sum of components. In general, the fluorescence will be dominated by a single component, which was the reference amplicon. This approach maximized the data extraction by also simultaneously generating information on the complexity of the sample as a function of the cycle number. The Maximum Entropy Method[1] can be used to determine the complexity of the sample. The Maximum Entropy Method leverages the entropy of the signal (the noise) by using it as a regularizing functional to constrain the solution and give the simplest possible (in the sense of the amount of contained information) compatibility with the data.

Appendix A - variance of equation (16) was

$$\alpha_\sigma^2=\left(\frac{\partial\sigma}{\partial n_a}\right)^2\alpha_{n_a}^2+\left(\frac{\partial\sigma}{\partial n_b}\right)^2\alpha_{n_b}^2+\left(\frac{\partial\sigma}{\partial n_c}\right)^2\alpha_{n_c}^2+\left(\frac{\partial\sigma}{\partial n_d}\right)^2\alpha_{n_d}^2+\left(\frac{\partial\sigma}{\partial\beta_2}\right)^2\alpha_{\beta_2}^2 \quad (A1)$$

Where

$$\left(\frac{\partial\sigma}{\partial n_a}\right)=-\left(\frac{1}{n_a}\right)^2\left[n_c n_b\ln\beta_2-n_c\ln\left(L_1/L_2\right)\right]\bullet\aleph \quad (A2)$$

$$\left(\frac{\partial\sigma}{\partial n_b}\right)=\left(\frac{n_c}{n_a}\ln\beta_2\right)\bullet\aleph \qquad \text{(A3)}$$

$$\left(\frac{\partial\sigma}{\partial n_c}\right)=\left[\frac{n_b}{n_a}\ln\beta_2-\frac{1}{n_a}\ln\left(L_1/L_2\right)\right]\bullet\aleph \qquad \text{(A4)}$$

$$\left(\frac{\partial\sigma}{\partial n_d}\right)=\left(-\ln\beta_2\right)\bullet\aleph \text{ (A5)}$$

$$\left(\frac{\partial\sigma}{\partial\beta_2}\right)=\frac{1}{\beta_2}\left(\frac{n_c n_b}{n_a}-n_d\right)\bullet\aleph \text{ (A6)}$$

And

$$\aleph=\exp\left[\ln\left(\frac{L_1}{L_2}\right)+\frac{n_c}{n_a}\left[\left(n_b-\frac{n_d n_a}{n_c}\right)\ln\beta_2-\ln\left(L_1/L_2\right)\right]\right] \qquad \text{(A7)}$$

6.6 Systematic* in vivo *identification of regulatory sequences within human gene domains

In the following example an embodiment of the invention in which methods based on gene expression profiles for prediction of drug induced liver damage are described. The example was presented by way of illustration of the present invention, and was not intended to limit the present invention in any way.

Understanding the human genome and those of other complex organisms will require comprehensive delineation of the functional elements that regulate transcription and other chromosomal processes. Actualization of the regulatory programs encoded in the genome sequence requires the interplay of chromatin proteins, *trans*-regulatory factors, and *cis*-active sequences. Packaging of a large, complex genome into the living nucleus was effected by histones and other ubiquitous chromatin proteins (Felsenfeld G. Chromatin unfolds. *Cell* 86,13-9 (1996);

Felsenfeld, G. & Groudine, M. Controlling the double helix. *Nature* 421, 448-53 (2003)). The result was the adoption of a highly compacted structure that eclipses the primary genome sequence and was hence highly repressive of genomic activity.

In vivo, regulatory elements and other *cis*-active modalities are found to coincide with focal alterations in chromatin structure (Gross, D.S. and Garrard, W. T. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57, 159-197 (1988); Elgin, S. C. Anatomy of hypersensitive sites. Nature 309, 213-4 (1984)). Such active genomic foci are detectable experimentally on the basis of pronounced sensitivity to cleavage when intact nuclei are exposed to DNA modifying agents, canonically the non-specific endonuclease DNaseI (Wu C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286, 854-60 (1980)). The co-localization of DNaseI Hypersensitive Sites (HSs) with cis-active elements spans the spectrum of known transcriptional and chromosomal regulatory activities including transcriptional enhancers, promoters, and silencers, insulators, locus control regions, and domain boundary elements (Felsenfeld G. Chromatin unfolds. Cell 86,13-9 (1996); Gross, D.S. and Garrard, W. T. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57, 159-197 (1988); Burgess-Beusse B, Farrell C, Gaszner M, Litt M, Mutskov V, et al. The insulation of genes from external enhancers and silencing chromatin. Proc. Natl Acad. Sci. U S A 99, 16433-7 (2002)). HSs have also been observed to coincide with sequences governing fundamental genomic processes including attachment to the nuclear matrix (Jarman AP, Higgs DR. Nuclear scaffold attachment sites in the human globin gene complexes. EMBO J. 7, 3337-44 (1988); Kieffer LJ, Greally JM, Landres I, Nag S, Nakajima Y et al. Identification of a candidate regulatory region in the human CD8 gene complex by colocalization of DNase I hypersensitive sites and matrix attachment regions which bind SATB1 and GATA-3. J. Immunol. 168, 3915-22 (2002)), and recombination Zhang Y, Strissel P, Strick R, Chen J, Nucifora G, et al. Genomic DNA breakpoints in AML1/RUNX1 and ETO cluster with topoisomerase II DNA cleavage and DNase I hypersensitive sites in t(8;21) leukemia. Proc. Natl Acad. Sci. U S A 99, 3070-5 (2002)). At present, it was therefore possible to generalize that any regulatory process which relies on the interaction of regulatory factors with specific genomic sequences will give rise to HSs (Felsenfeld G. Chromatin unfolds. Cell 86,13-9 (1996); Gross, D.S. and Garrard, W. T. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57, 159-197 (1988)).

This example exploits the potential of chromatin accessibility for generic in vivo identification of regulatory sequences within a genomic locus of interest. Localization of DNaseI hypersensitive sites in the context of the native genome had traditionally relied on the Southern transfer/indirect end-labeling approach (Wu C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature 286, 854-60 (1980)). While widely applied, this technique was only semi-quantitative and suffers from numerous technical and resource-related limitations that prevent its application on a genomic scale. Thus, high-resolution, quantitative in vivo profiles of DNaseI sensitivity could be obtained by spanning a genomic locus with a series of contiguous amplicons and measuring the relative digestion sensitivity of each amplicon using real-time quantitative PCR. In such profiles, hypersensitive sites appear as outliers that are detectable using a statistical algorithm. High-resolution chromatin profiles would thus readily resolve foci of DNaseI hypersensitivity at the sequence level, and hence permit localization of the regulatory sequences active within the study region and tissue. Preliminary experiments had shown the feasibility of quantitative analysis of selected known HSs with real-time PCR (McArthur, M., Gerum, S. & Stamatoyannopoulos, G. Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse beta-globin LCR. J. Mol. Biol.313, 27-34 (2001)). The scalability of quantitative PCR allows high-throughput quantitative chromatin profiling (QCP) which was capable of de novo identification of DNaseI hypersensitive sites on a genomic scale.

Quantitative chromatin profiles rapidly delineate nuclease hypersensitive sites

QCP was tested with the hypersensitive sites of the human beta-globin locus control region (LCR) as a model system. The human beta-globin LCR comprises an array of functional elements coinciding with five major DNaseI hypersensitive sites (designated HS1-HS5 in 3'→5' order) located upstream of the epsilon-globin gene on chromosome 11 (Tuan D, Solomon W, Li Q, London IM. The "beta-like-globin" gene domain in human erythroid cells. Proc. Natl Acad. Sci. U S A 82, 6384-8 (1985); Forrester WC, Thompson C, Elder JT, Groudine M. A developmentally stable chromatin structure in the human beta-globin gene cluster. Proc. Natl Acad. Sci. U S A 83, 1359-63 (1986)). These sites had been localized precisely at the sequence level through extensive functional and chromatin structural studies (Forrester WC, Takegawa S, Papayannopoulou T, Stamatoyannopoulos G, Groudine M. Evidence for a locus activation region: the formation of developmentally stable

hypersensitive sites in globin-expressing hybrids. Nucleic Acids Res. 15, 10159-77 (1987); Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. Cell 51, 975-85 (1987); Stamatoyannopoulos, G. and Grosveld, F. Hemoglobin Switching. In Stamatoyannopoulos, G, Majerus, P., Perlmutter, R., Varmus, H. The molecular basis of blood diseases (W.B. Saunders, Philadelphia, 2001)).

To produce a quantitative chromatin profile of the human beta-globin LCR, an erythroid cell line (K562) was used, in which the lineage-specific HSs of the LCR are known to be active (Tuan D, Solomon W, Li Q, London IM. The "beta-like-globin" gene domain in human erythroid cells. Proc. Natl Acad. Sci. U S A 82, 6384-8 (1985)). A series of 89 contiguous amplicons (mean length 225bp) were designed across a ~20kb interval spanning HS1-5. Purified genomic DNA was prepared from DNaseI-treated and untreated K562 chromatin samples. The sensitivity of each amplicon to DNaseI digestion was measured by quantifying relative copy ratios between DNase-treated and untreated samples. In order to arrive with a common DNase sensitivity scale that could be applied to other loci, relative copy ratios were normalized to a standardized reference amplicon from the Rhodopsin gene locus on chromosome 3. The Rhodopsin gene was transcriptionally inactive and DNaseI-resistant in all cell types employed herein (data not shown). Nine replicate measurements were performed for each amplicon and a relative DNaseI sensitivity profile was constructed (Fig. 14a).

By definition, hypersensitive sites should appear as statistical outliers within a quantitative profile of DNaseI sensitivity. To identify HSs within the profile, a rigorous statistical approach was employed. First, the trend or 'baseline' behavior of DNaseI sensitivity across the locus was determined. Next, measurement errors for DNase sensitivity values clustered around the baseline, and hence confidence bounds on outliers and extreme values for this distribution, were determined (Fig. 14a). Outliers that displayed clustering behavior (low variance) under repeated measurement with respect to the mean measurement error were identified. In a final step, a signal-to-noise ratio (SNR) was assigned to each outlier to quantify the significance of its observation relative to the baseline. The SNR was a broadly applied instrument in quantitative assays and values >3 were regarded as statistically significant. Computed hypersensitivity SNRs across the beta-globin LCR were shown in Fig. 14b. If the region of *in vivo* DNaseI hypersensitivity extends beyond

the span of a given amplicon, then two or more contiguous amplicons will be expected to display hypersensitivity. This was manifested in a wide single peak.

Using QCP, all of the DNaseI hypersensitive sites of the human beta-globin LCR were correctly and rigorously identified (HS1 to HS5; Fig. 14b). A minor hypersensitive site described previously (Forrester WC, Takegawa S, Papayannopoulou T, Stamatoyannopoulos G, Groudine M. Evidence for a locus activation region: the formation of developmentally stable hypersensitive sites in globin-expressing hybrids. Nucleic Acids Res. 15, 10159-77 (1987) in K562 cells and designated HS-7.2 (to reflect its position 7.2kb upstream of the epsilon-globin cap site) was also found. HS1;HS2, and HS5 comprised more than one contiguous hypersensitive amplicon. In each case, the amplicon with the highest SNR was found to correspond precisely with sequences that had been determined previously to encompass the core regulatory factor binding domains (Stamatoyannopoulos, G. and Grosveld, F. Hemoglobin Switching. In Stamatoyannopoulos, G, Majerus, P., Perlmutter, R., Varmus, H. The molecular basis of blood diseases (W.B. Saunders, Philadelphia, 2001)) (Fig. 14c).

High-throughout mapping of complex human gene regulatory regions

Chromatin profiles for the alpha-globin upstream regulatory region on chromosome 16, the adenosine deaminase (ADA) locus on chromosome 20, the CD2 locus on chromosome 1, and the T-cell receptor-alpha downstream regulatory region on chromosome 14 were also produced. For the alpha-globin locus profile, the same K562 samples employed for the study of the beta-globin LCR, and in which the alpha-globin locus was known to be active (Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J, Pretorius IM, Ayyub H. A major positive regulatory region located far upstream of the human alpha-globin gene locus. Genes Dev. 4, 1588-601 (1990)), were employed. The ADA, CD2, and TCR-alpha profiles were produced using unstimulated Jurkat cells.

A chromatin profile spanning 66.4kb upstream of the zeta-globin gene was obtained (Fig. 15a). This region contains the alpha-globin major regulatory element, which was situated in an intron of an unrelated constitutively active upstream gene (Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J, Pretorius IM, Ayyub H. A major positive regulatory region located far upstream of the human alpha-globin gene locus. Genes Dev. 4, 1588-601 (1990)). The profile revealed eight hypersensitive

sites at –7.8kb, -9.8kb, -14.2kb, -32.9kb, -39.4kb, -45.6kb, -47.7kb, and –55.8kb relative to the zeta-globin cap site (Fig. 15a). These features coincided with all major HSs previously reported following lengthy and exhaustive studies of the region with conventional hypersensitivity assays in this tissue (Stamatoyannopoulos, G. and Grosveld, F. Hemoglobin Switching. In Stamatoyannopoulos, G, Majerus, P., Perlmutter, R., Varmus, H. The molecular basis of blood diseases (W.B. Saunders, Philadelphia, 2001)). Notably, the HSs identified in the profile include all of the major transcriptional control elements of the alpha-globin upstream regulatory domain (Higgs DR, Wood WG, Jarman AP, Sharpe J, Lida J, Pretorius IM, Ayyub H. A major positive regulatory region located far upstream of the human alpha-globin gene locus. Genes Dev. 4, 1588-601 (1990); Vyas P, Vickers MA, Simmons DL, Ayyub H, Craddock CF, Higgs DR. Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin. Cell 69, 781-93 (1992); Sharpe JA, Wells DJ, Whitelaw E, Vyas P, Higgs DR, Wood WG. Analysis of the human alpha-globin gene cluster in transgenic mice. Proc. Natl Acad. Sci. U S A 90, 11262-6 (1993)).

The ADA gene was under control of a regulatory array positioned centrally within the 18kb first intron (Aronow B., Lattier D., Silbiger R., Dusing M., Hutton J. et al. Evidence for a complex regulatory array in the first intron of the human adenosine deaminase gene. Genes Dev. 3, 1384-400 (1989)). This array was distinguished by the presence of a strong central tissue-specific DNaseI hypersensitive site (designated HSIII). In thymus and cell lines with a T-cell phenotype, ^{ADA}HSIII was the only strongly evident site in this region and encodes a powerful transcriptional enhancer which was the dominant regulatory element of the ADA gene (Aronow, B.J., Silbiger, R.N., Dusing, M.R., Stock, J.L., Yager, K.L., et al. Functional analysis of the human adenosine deaminase gene thymic regulatory region and its ability to generate position-independent transgene expression. *Mol. Cell. Biol.* **12**, 4170-4185 (1992)). A chromatin profile spanning 58.4kb of the ADA gene locus in Jurkat cells (Fig. 15b) was obtained. The central hypersensitive site coincides precisely with ^{ADA}HSIII, the major regulatory element of the ADA LCR. Three additional statistically significant events were detected. One site (^{ADA}HSA) was present over the ADA proximal promoter region. A second site (^{ADA}HSB) at the 3' end of the locus occurs in the first intron of the downstream protein kinase inhibitor G gene. A third site (^{ADA}HSC) occurs 5' of an upstream gene.

The CD2 gene was controlled by a strong enhancer element positioned 3' to the gene and marked by a major DNaseI HS (^{CD2}HS3) (Greaves DR, Wilson FD, Lang G, Kioussis D. Human CD2 3'-flanking sequences confer high-level, T cell-specific, position-independent gene expression in transgenic mice. *Cell* **56**, 979-86 (1989)). This element was critical for high-level expression of CD2 and for its tissue and lineage specificity (Lake RA, Wotton D, Owen MJ. A 3' transcriptional enhancer regulates tissue-specific expression of the human CD2 gene. *EMBO J.* **9**, 3129-36 (1990)). ^{CD2}HS3 also exhibits locus control region activity in transgenic systems (Lang G, Mamalaki C, Greenberg D, Yannoutsos N, Kioussis D. Deletion analysis of the human CD2 gene locus control region in transgenic mice. *Nucleic Acids Res.* 19, 5851-6 (1991)). A hypersensitive site had also been defined over the transcriptional promoter (^{CD2}HS2); this site had been described previously only in CD2-positive cells (Wotton D, Flanagan BF, Owen MJ. Chromatin configuration of the human CD2 gene locus during T-cell development. *Proc. Natl Acad. Sci. U S A* 86, 4195-9 (1989)). By contrast, ^{CD2}HS3 was present in all cells of the T-cell lineage, including progenitor cells in which the CD2 gene was not yet transcriptionally active. The CD2 genic regions (exons/introns) have not been explored previously for the presence of HSs. QCP was applied to a 26.7kb region of the CD2 locus in Jurkat cells encompassing the CD2 gene and its flanking sequences (Fig. 15c). The CD2 profile revealed 3 hypersensitive sites. The 5'-most site corresponds to the previously described ^{CD2}HS2 promoter site. The prominent 3'-most site corresponds to the CD2 major regulatory element encoded by ^{CD2}HS3. The analysis also revealed a novel site of equal prominence with ^{CD2}HS2 and located within the second intron.

To test the ability of QCP to delineate complex promoter elements, the c-*myc* gene on chromosome 8 was studied. c-*myc* belongs to a class of highly regulated genes for which multiple distinct promoter elements had been described. 3 sites of transcription complex (proximal promoter) formation had been defined for the c-*myc* gene, designated P0, P1, and P2 (Bentley DL, Groudine M. Novel promoter upstream of the human c-myc gene and regulation of c-myc expression in B-cell lymphomas. *Mol. Cell. Biol.* **6**, 3481-9 (1986)). Transcription from different promoters may occur within the same cell type or under different growth or differentiation conditions (Dyson PJ, Littlewood TD, Forster A, Rabbitts TH. Chromatin structure of transcriptionally active and inactive human c-myc alleles. *EMBO J.* **4**, 2885-91 (1985)). Each was marked by the presence of a 5'-apposed DNaseI-hypersensitive

site, designated HSI-III, respectively. In cells expressing c-*myc*, all three sites were generally observed. Additionally, the surrounding regions had been surveyed for the presence of HSs in various cell types, with description of a number of major sites that appear to form in a tissue-specific pattern (Mautner J, Joos S, Werner T, Eick D, Bornkamm GW, Polack A. Identification of two enhancer elements downstream of the human c-myc gene. *Nucleic Acids Res.* **23**, 72-80 (1995)).

A 30.4kb chromatin profile of the c-*myc* locus in HepG2 hepatocellular carcinoma cells was obtained. Expression profiling had revealed the c-*myc* gene to be transcriptionally active. The profile revealed 5 distinct HSs (Fig. 15d). 3 central sites corresponded to the previously-described ^{myc}HSI-III. As such, the chromatin profile correctly delineated the structure of the c-*myc* promoter. Hypersensitive sites located ~11kb upstream of the P1 promoter and ~1.5kb downstream of the gene were also observed; these correspond with sites previously documented in intestinal Colo320 cells (Mautner J, Joos S, Werner T, Eick D, Bornkamm GW, Polack A. Identification of two enhancer elements downstream of the human c-myc gene. *Nucleic Acids Res.* **23**, 72-80 (1995)) and designated HS5' and HS1.5 (respectively).

Comprehensive delineation of regulatory elements across a multi-gene locus

The entire human beta-like globin gene domain on chromosome 11 was also profiled. This domain comprises 5 genes (ε, $^{G}\gamma$, $^{A}\gamma$, δ, and β) organized in a 5' to 3' fashion that corresponds to their timing of expression during development and differentiation (Stamatoyannopoulos, G. and Grosveld, F. Hemoglobin Switching. In Stamatoyannopoulos, G, Majerus, P., Perlmutter, R., Varmus, H. The molecular basis of blood diseases (W.B. Saunders, Philadelphia, 2001)). Like its LCR described above, the beta-like gene locus had also been extensively characterized at the chromatin level. In addition to the five major DNaseI hypersensitive sites that constitute the LCR, the promoters of the ε-, $^{G}\gamma$, $^{A}\gamma$, δ-, and β-globin genes have all been shown to be DNase-hypersensitive in erythroid cells.

The 90.4kb beta-globin chromatin profile revealed 9 strong hypersensitive sites and 14 intermediate- to low-prominence sites. Significantly, this feature set encompassed all previously-identified major HSs in this tissue (Fig. 16). QCP enabled direct quantitative comparison between these elements in the same tissue

sample, revealing that the ε-, $^{G}\gamma$-, $^{A}\gamma$-, and δ-globin promoters exhibit comparable hypersensitivity to the major HSs in the LCR.

Cis-active sequences were components of higher-order chromatin structures

Since QCP enabled essentially simultaneous study of an entire locus, it was possible to contextualize HS elements further on a quantitative basis relative to one another, to their immediate flanking regions, and to their chromosomal domains generally. Previous studies have suggested that sequences flanking certain elements within the beta-globin and ADA LCRs contribute to their activity *in vivo*. The chromatin profiles revealed the presence of numerous prominent perturbations representing zones of significantly increased sensitivity extending over 1-3 kilobases (Fig. 17). Each of these regions was associated with an HS, typically of higher signal strength. Interestingly, the structural configuration of each region appeared to be characteristic and was highly reproducible. One explanation for these observations was that local disruptions at the site of cooperative regulatory factor binding were propagated due to sequence-specific loss of histone interactions. Alternatively, such elements may act as foci for the recruitment of chromatin remodeling activities such as histone RStylases which exert their effects on the neighboring regions resulting in local relaxation of the chromatin fiber. A third (inclusive) possibility was that the effect was an intrinsic functional property of the flanking sequences, but requires formation of a hypersensitive site in order to become manifest.

Chromatin analysis and comparative genomics

Comparative genomic analyses represent a conceptually attractive approach for identification of regulatory sequences (Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* 4, 251-62 (2003)). The central hypothesis of such studies was that functionally important sequences will exhibit selective pressures that propagate over evolutionary distances (Dermitzakis ET., Reymond A., Lyle R., Scamuffa N., Ucla C. et al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* 420, 578-82 (2002)). The comparative genomics of nuclease hypersensitive sites had not been formally evaluated. While it was clear that certain hypersensitive regulatory elements had been highly conserved during vertebrate and

particularly mammalian evolution (Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, et al. Distinguishing regulatory DNA from neutral sites. *Genome Res.* 13, 64-72 (2003)), it was also evident that many such elements exhibit little or no selective conservation above local background (Flint J, Tufarelli C, Peden J, Clark K, Daniels RJ, et al. Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the alpha globin cluster. *Hum. Mol. Genet.* 10, 371-82 (2001)).

A priori, quantitative chromatin profiles and comparative studies should be highly complementary. To examine this, the human T-cell receptor-alpha (TCR-alpha) locus was studied. Knowledge of the regulatory structure of this locus had heretofore been inferred indirectly from functional studies of the homologous murine locus (Hong NA, Cado D, Mitchell J, Ortiz BD, Hsieh SN et al. A targeted mutation at the T-cell receptor alpha/delta locus impairs T-cell development and reveals the presence of the nearby antiapoptosis gene Dad1. *Mol. Cell. Biol.* 17, 2151-7 (1997)), including the identification of hypersensitive sites and a locus control region (Diaz, P., D. Cado, and A. Winoto. A locus control region in the T cell receptor a/d locus. *Immunity* 1, 207-217 (1994)). The TCR-alpha locus was embedded in a large, highly-conserved segment of human chromosome 14 (Koop BF, Hood L. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7, 48-53 (1994)). In mouse, the TCR-alpha regulatory domain lies 3' to the gene and extends for at least 20 kilobases to the ubiquitously-expressed downstream *Dad1* gene. Two regulatory regions had been delineated downstream of the murine TCR-alpha gene (Diaz, P., D. Cado, and A. Winoto. A locus control region in the T cell receptor a/d locus. *Immunity* 1, 207-217 (1994)). These regions coincide with two major HSs, designated HS1 and HS6. ^{TCR}HS1 had been reported to comprise two closely juxtaposed sites, with the 3' site designated HS1' (Ortiz, B.D., Cado, D., and Winoto, A. A new element within the T-cell receptor alpha locus required for tissue-specific locus control region activity. *Mol. Cell. Biol.* 19, 1901-1909 (1999)). Several minor sites had been reported to lie between ^{muTCR}HS1 and ^{muTCR}HS6 but it was unclear whether they represent true hypersensitive sites as their intensity does not vary as expected with increasing DNaseI treatment (Diaz, P., D. Cado, and A. Winoto. A locus control region in the T cell receptor a/d locus. *Immunity* 1, 207-217 (1994)); also, this intervening region does not appear to contribute to TCR-alpha regulation (Ortiz, B.D., Cado, D., and Winoto, A. A new element within the T-cell

receptor alpha locus required for tissue-specific locus control region activity. *Mol. Cell. Biol.* 19, 1901-1909 (1999)).

To delineate the regulatory structure of the homologous human domain, a 26.3kb downstream of the human TCR-alpha gene in Jurkat T-cells was profiled (Fig. 18a). Four prominent HSs were detected. Alignment of syntenic regions of the mouse and human TCR-alpha sequences using rVista (Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12, 832-9 (2002)) revealed extensive sequence conservation along the TCR-alpha locus (Fig. 18b). Three of the identified human HSs overlie conserved sequences that correspond with the murine ^{TCR}HS1, ^{TCR}HS1′, and ^{TCR}HS6. However, it was not possible to distinguish these sequences and that underlying the prominent HS in the first intron of *Dad1* from many others in the locus that exhibit similar levels of evolutionary conservation. Similar analyses were performed for the other loci examined with the same result. As such, QCP provided powerful functional illumination of comparative genomic data. Several features thus serve to distinguish chromatin profiles from comparative genomic analyses. Firstly, chromatin studies provide direct *in vivo* functional information for human genome sequences in particular tissue regulatory environments. Secondly, chromatin analysis can be applied generically across the genome, irrespective of regional evolutionary rates. Thirdly, chromatin profiling was extensible to any eukaryotic organism for which genomic sequence was extant (whether partial or complete), irrespective of the availability of sequence from homologous organisms.

Advantages and limitations of chromatin profiling

High-throughput, high-resolution mapping of chromatin structure using quantitative PCR had numerous advantages over conventional DNaseI hypersensitivity assays. QCP was extremely rapid and can be performed over tens or even many hundreds of kilobases per day (with a multi-machine set-up). QCP was not dependent on the availability of convenient restriction sites or high-specificity probes, nor on any of the many operator-dependent parameters that render conventional hypersensitivity assays time-consuming and technically challenging. Unlike conventional chromatin assays, QCP was highly resource-efficient and can be performed using very small quantities of tissue (such that a single typical tissue

preparation could yield sufficient material for thousands of assays). It provides quantitative, highly reproducible data that enable determination of hypersensitivity directly in a sequence-specific fashion, and it was applicable to any genomic locus.

Another advantageous feature of QCP was that localization of *cis*-active sequences was generic: a broad range of elements may be localized in a single profile including transcriptional enhancers, promoters, locus control regions, and domain boundary elements. In cases where data were available, peaks of hypersensitivity corresponded precisely with previously described core transcription factor binding domains.

The profiling resolution was restricted by amplicon design parameters. For survey applications, it might be desirable to increase the average amplicon size and thereby increase the throughput and cost-efficiency of the assay. It had been determined that lowering the profiling resolution by increasing the amplicon size beyond a mean of 250bp results in a precipitous erosion of signal-to-noise ratio and hence the ability to detect HSs (data not shown). Theoretically, the method had a lower amplicon size limit of approximately 50-70bp. However, in practice no significant improvement in detection of HS regions was observed with smaller amplicons (data not shown). This was not unexpected as our data revealed that the hypersensitivity phenomenon was frequently distributed over >250bp intervals. In addition, decreasing amplicon size resulted in a significant deterioration in sequence coverage due to the lack of flexibility in selecting appropriate primer pairs.

Application to genetic analysis

Approximately 98% of genetic variation was found in non-coding regions (Kruglyak L, Nickerson DA. Variation was the spice of life. *Nat Genet.* **27,** 234-6 (2001)); only a tiny fraction of these variants occur within functionally important regions. As such, the ability to discriminate systematically SNPs and other variants that occur within *cis*-active sequences (and hence were of potential functional consequence) should have a significant impact on our ability to ascertain genetic causes of disease. The relative contribution of functional regulatory vs. coding variants to disease was unknown. However, it was expected that *cis*-regulatory variation will form a major component of the genetic basis of quantitative traits, including those involved in many common diseases. Combining QCP with studies of

allelic variation in functional elements should greatly facilitate large-scale identification of heretofore-elusive regulatory variants.

Methods

Cell Culture. K562 (ATCC), Jurkat (JCRB) and HepG2 (ATCC) were cultured in humidified incubators at 37° C and 5% CO_2 in air. K562 and Jurkat cells were grown in RPMI (Invitrogen, Carlsbad, CA) supplemented with 10% FBS. HepG2 cells were cultured in MEM (Invitrogen, Carlsbad, CA) with 10% FBS. Suspension cultures were harvested for nuclei preps at a density of $5x10^5$ cells/ml. HepG2 was harvested at 80% confluency at a cell density of $2x10^5$ cells/ml. Accutase (Innovative Cell Technologies, San Diego, CA) was used to detach adherent cells.

DNaseI digestion and DNA purification. DNaseI digestions were carried out according to a standard protocol (Reitman, M., Lee, E., Westphal, H. & Felsenfeld, G. An enhancer/locus control region was not sufficient to open chromatin. Mol. Cell. Biol. 13, 3990-8 (1993)). Following DNaseI treatments, DNA was purified using the Puregene system (Gentra Systems, Minneapolis, MN) according to the manufacturer's protocol and resuspended in 10mM Tris-Cl, pH 8.0. Samples were quantitated in triplicate using a Spectramax 384 Plus UV spectrophotometer (Molecular Devices Corporation, Sunnyvale, CA).

Primer selection. Primers were designed to amplify contiguous or minimally overlapping ~250 base amplicons across target genomic regions. Primers were designed using Primer3 (Rozen, S., Skaletsky, H.J. Primer3 on the WWW for general users and for biologist programmers. *In:* Krawetz, S. and Misener, S. *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (Humana Press, NJ, 2000)) restricting several parameters, including target amplimer size (250 bp +/- 50 bases); primer T_m (optimal, 60°C +/- 2°C); %GC (50% optimal, range 40-80%), and length (optimal 24, range 19-27); and the poly X (maximum 4). Primers were then scanned for repetitive sequences by BLAST alignment with the *Alu* and NR databases.

Quantitative PCR. 15μL real-time quantitative PCR reactions were assembled using 0.9μM forward and reverse primers, 30 ng template DNA (untreated or DNaseI-treated) and master mix composed of 1X FastStart buffer (Roche), 200μM of each dATP, dCTP, dGTP, dTTP, 3mM $MgCl_2$ and FastStart Taq DNA polymerase (0.033 U/μL). The reaction mixture was supplemented with 0.33X SYBR green I stain and

300 nM 6-ROX (Molecular Probes, Eugene, OR) to detect the accumulation of PCR product during amplification and normalize fluorescence intensity, respectively. All qPCR reactions were set up robotically with a Biomek FX (Beckman, Fullerton, CA). Samples were run in triplicate on individual 384-well plates, and thermalcycled with an ABI 7900HT Sequence Detection System (Applied Biosystems, Foster City, CA).

Primary Data Analysis. Measured data were processed in four phases: 1) C_t determination, 2) amplification efficiency correction, 3) melting curve analysis, and 4) calculation of DNaseI sensitivity ratios. Normalized fluorescence data were exported using the ABI SDS software (v2.0). An amplification curve and Nth-order polynomial fit was then computed for each reaction. Cycle threshold (C_t) values were then determined for each curve. The amplification efficiency (Liu W, Saint DA. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. *Anal. Biochem.* 302, 52-9 (2002)) of a reference amplicon selected from the inactive and DNase-insensitive Rhodopsin locus (3q21-q24) was determined empirically for every reaction plate using a standard dilution series of DNA and the equation $E = 10^{-1/slope}$. We then derived the efficiency of each test amplicon was from the slope of the linear region of the amplification curve (Ramakers C, Ruijter JM, Deprez RH, Moorman AF. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci Lett.* 339, 62-6 (2003)). Efficiency corrections were then performed on all test amplicons with respect to the reference amplicon, following which we calculated relative copy number differences using the comparative C_t method (Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402-8 (2001)). Melting curve analysis was conducted for each amplicon to discard those yielding multiple products. Efficiency-corrected C_t values were then used to compute a relative copy number ratio by applying the formula $2^{-\Delta\Delta Ct}$ or $2^{-[\text{treated (target – reference)} - \text{calibrator (target-reference)}]}$. Relative DNaseI sensitivity ratios (=relative copy ratios) were thus obtained. Ratios < 1 were indicative of relative copy loss due to preferential cleavage of chromatin by DNaseI.

Statistical analysis of DNaseI sensitivity and hypersensitivity To determine the baseline, a linear pass through the locus dataset using a 20% trimmed mean was performed to remove egregious outliers. The remaining data were then processed using a LOWESS (Locally Weighted Least Squares) smoother (Chambers, J.M.,

Cleveland, W.S., Tukey, P.A. *Graphical Methods for Data Analysis* (Wadsworth, CA, 1983)) that was adapted for robust locally-weighted time series and scatter plot smoothing and was particularly suited for non-Gaussian values. Data were mean-centered about the moving baseline and outliers of this distribution were determined using the median average deviation approach (Rousseeuw, P.J., van Zomeren, B.C. Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* 85, 633-639 (1990)), where a value X was declared to be an outlier if $0.6745 \mid X - M \mid / MAD > 2.24$ where M was the median and MAD was the average median deviation. Hypersensitive sites were then identified as clusters of DNaseI sensitivity scores over the same genomic position whose 20% trimmed mean lies strictly below the interpolated value at the lower shifted baseline. To determine the signal-to-noise ratio (SNR) we computed $S/N_i = \mid HS_i - B_i \mid / MAD_B (\sigma_c / \sigma_{HS})^2$ where S/N_i, the signal-to-noise ratio at site i was measured as the absolute deviation of the 20% trimmed mean of the HS cluster from the interpolated baseline, divided by the median average deviation of the centered baseline. The remaining term $(\sigma_c / \sigma_{HS})^2$ was a small correction factor that penalizes larger variances in HS clusters and rewards highly compact clusters that were strongly indicative of HS sites. It was simply the ratio of the variance of the data comprising the HS cluster to the average variance of data assigned to HS clusters computed over all scored data.

7. REFERENCES CITED

All references cited herein were incorporated herein by reference in their entirety and for all purposes to the same extent as if each individual publication or patent or patent application was specifically and individually indicated to be incorporated by reference in its entirety for all purposes.

Many modifications and variations of the present invention can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. The specific embodiments described herein were offered by way of example only, and the invention was to be limited only by the terms of the appended claims along with the full scope of equivalents to which such claims were entitled.